2

Etapas da CRISP-DM:

1. **Compreensão do Negócio:** Identificar objetivos do projeto e converter em metas específicas de mineração de dados.
2. **Compreensão dos Dados:** Coleta, descrição e análise inicial dos dados.
3. **Preparação dos Dados:** Limpeza, integração e transformação dos dados para o formato adequado.
4. **Modelagem:** Seleção e aplicação de técnicas de modelagem de dados.
5. **Avaliação:** Verificar se o modelo atende aos objetivos estabelecidos.
6. **Implementação:** Aplicar os modelos em ambiente operacional para uso contínuo.

1

- **Aprendizagem supervisionada:** Envolve dados rotulados, onde o modelo aprende a mapear entradas para saídas específicas. Exemplos de técnicas:
  1. **Regressão Linear:** Previsão de preços de casas.
  2. **Árvores de Decisão:** Classificação de e-mails como spam ou não spam.
- **Aprendizagem não supervisionada:** Trabalha com dados não rotulados para encontrar padrões ou estruturas. Exemplos de técnicas:
  1. **K-Means Clustering:** Segmentação de clientes.
  2. **Análise de Componentes Principais (PCA):** Redução de dimensionalidade em dados genômicos.
- **Aprendizagem por reforço:** Baseia-se em agentes que aprendem interagindo com um ambiente, recebendo recompensas ou punições. Exemplos de técnicas:
  1. **Q-Learning:** Jogos como xadrez ou Go.
  2. **Deep Q-Networks (DQN):** Controle de robôs em ambientes dinâmicos.

N° _____ CURSO _____

NOME eduardovski

_____

GRUPO 1
(4 valores)

RESPONDA ÀS QU

QUESTÃO 1

No desenvolvimento de sistemas de aprendizagem automática (*machine learning*) podem ser utilizados diferentes paradigmas de aprendizagem.

Neste contexto pretende-se que:

a) caracterize os paradigmas de aprendizagem supervisionada, não supervisionada e por reforço;

b) apresente dois exemplos de técnicas de cada paradigma, ilustrando-os com casos de aplicação.

QUESTÃO 2

O processo de desenvolvimento de uma solução de aprendizagem automática envolve diversas etapas, que podem diferir de acordo com a metodologia escolhida.

Tendo em consideração a metodologia CRISP-DM, pretende-se que enumere e descreva as suas etapas.

GRUPO 2
(4 valores)

Responda às questões deste grupo no espaço reservado PREENCHENDO OS ESPAÇOS VAZIOS com as expressões devidas de modo que a afirmação seja correta.

QUESTÃO 1

No contexto da utilização de técnicas de aprendizagem automática (*machine learning*), a adoção de uma metodologia para a extração de conhecimento descreve e cria _____etapas_____ pelos quais deverá passar o desenvolvimento de um projeto de extração de conhecimento para _____a tomada de decisao_____.

QUESTÃO 2

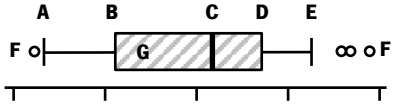
A metodologia de extração de conhecimento que se desenvolve em 5 etapas, a saber, _____sample / amostra_____, _____Explore_____, _____Modify_____, _____Model_____ e _____Acess / avaliacao_____, denomina-se SEMMA.

QUESTÃO 3

Máquina de Vetores de Suporte (*Support Vector Machine*) é uma técnica _____supervisionada_____ de aprendizagem automática que pode ser utilizada para resolver problemas de _____regressao_____ e de _____classificacao_____.

QUESTÃO 4

Num diagrama de caixa (*boxplot*), como no exemplo à direita, o ponto **C** corresponde à _____mediana_____, a caixa **G** representa _____intervalo interquartil_____ dos dados do estudo, e os círculos **F** identificam os valores _____outliers_____ do *dataset*.



N° _____

GRUPO 3
(6 valores)

RESPONDA ÀS QUESTÕES DESTE GRUPO NO ESPAÇO RESERVADO.

Considere o *dataset Titanic*, utilizado por diversas vezes ao longo do semestre. Considere também o excerto de código apresentado na Figura 1, onde é apresentada a construção e avaliação de um modelo de aprendizagem automática.

QUESTÃO 1

O excerto apresentado contém imprecisões. Identifique e corrija-as utilizando o espaço disponível ao lado da figura (não deve copiar todo o excerto, mas apenas aquilo que corrigiu).

```
[1]  df = pd.read_csv('titanic_dataset.csv')
[2]  X = df.drop(['Survived', 'Age', 'PassengerId', 'Name',
           'Ticket', 'Cabin', 'Embarked', 'Sex'], axis=1)
[3]  y = df['Survived']
[4]  sex_ohe = pd.merge(df['Sex'], drop_first=True)
[5]  embarked_ohe = pd.merge(df['Embarked'], drop_first=True)
[6]  X = pd.concat([X, sex_ohe, embarked_ohe], axis=1)
[7]  X_train, X_test, y_train, y_test =
                train_test_split(y, X, test_size=0.3)
[8]  model = Sequential()
[9]  model.add(Dense(16, input_dim=y.shape[1],
                                activation='relu'))
[10] model.add(Dense(8, activation='relu'))
[11] model.add(Dense(1, activation='sigmoid'))
[12] model.compile( loss = 'binary_crossentropy',
                    optimizer = 'adam',
                    metrics = ['mse'])
[13] model.transform(X_train, y_train, epochs=50,
                batch_size=32)
[14] loss, acc = model.evaluate(X_train, y_train)
```

python

```
[4] sex_ohe = pd.get_dummies(df['Sex'], drop_first=True)  # Substituir pd.mer
[5] embarked_ohe = pd.get_dummies(df['Embarked'], drop_first=True)  # Correçã
[7] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
[9] model.add(Dense(16, input_dim=X.shape[1], activation='relu'))  # Corrigir
[13] model.fit(X_train, y_train, epochs=50, batch_size=32)  # Substituir 'tra
```

Figura 1. Excerto de um modelo de aprendizagem.

QUESTÃO 2

Identifique a técnica de aprendizagem utilizada no excerto de código apresentado na Figura 1, e indique quatro hiperparâmetros passíveis de serem modificados para afinar o modelo.

Questão 2: Técnica de aprendizagem e hiperparâmetros

- **Técnica de aprendizagem:** A técnica usada é **Redes Neurais Artificiais (RNA)**, uma abordagem de aprendizagem supervisionada.
- **Quatro hiperparâmetros ajustáveis:**
  1. **Número de neurônios em cada camada:** Controla a capacidade de aprendizado da rede.
  2. **Taxa de aprendizado:** Define o passo dado durante a otimização.
  3. **Função de ativação:** Pode variar (e.g., ReLU, sigmoid, tanh) para diferentes comportamentos do modelo.
  4. **Número de épocas (epochs):** Define quantas vezes o modelo passa pelos dados durante o treinamento.

QUESTÃO 3

Admita que o *dataset Titanic* não está balanceado. Descreva de que forma este desbalanceamento influencia o modelo.

Questão 3: Impacto do desbalanceamento no modelo

Se o dataset Titanic não estiver balanceado, o impacto será:

1. **Previsões enviesadas:** O modelo pode aprender a prever principalmente a classe majoritária (por exemplo, "não sobreviveu"), ignorando os padrões da classe minoritária ("sobreviveu").
2. **Métricas distorcidas:** A acurácia pode parecer alta, mas métricas como **recall** e **F1-score** para a classe minoritária seriam baixas.
3. **Soluções para lidar com o desbalanceamento:**
   - **Reamostragem:** Balancear o dataset usando técnicas como *oversampling* (aumentar a classe minoritária) ou *undersampling* (reduzir a classe majoritária).
   - **Atribuir pesos às classes:** Ajustar a função de perda para penalizar erros na classe minoritária.
   - **Estratégias de dados sintéticos:** Usar métodos como **SMOTE (Synthetic Minority Oversampling Technique)** para gerar novos exemplos da classe minoritária.

N° _____

GRUPO 4
(6 valores)

Comente as afirmações seguintes, assinalando a sua veracidade (**V**) ou falsidade (**F**), justificando a resposta EXCLUSIVAMENTE no espaço disponibilizado.
NÃO SÃO CONSIDERADAS respostas para as quais não exista justificação expressa.

QUESTÃO 1 [v] No desenvolvimento de sistemas de aprendizagem automática, a fase de preparação de dados tem particular importância porque os dados obtidos do «mundo físico» são incompletos, contêm lixo e são falsos.

Verdadeiro (V)

- **Justificativa:** Os dados obtidos do mundo físico frequentemente apresentam problemas como valores incompletos, ruídos e inconsistências. A preparação de dados é essencial para corrigir essas questões e garantir a qualidade do modelo.

QUESTÃO 2 [f] Técnicas de aprendizagem automática baseadas no desenvolvimento de árvores de decisão são utilizadas exclusivamente para a resolução de problemas de classificação.

Falso (F)

- **Justificativa:** Técnicas baseadas em árvores de decisão, como *Decision Tree* ou *Random Forest*, podem ser usadas tanto para **classificação** quanto para **regressão**, dependendo do problema.

QUESTÃO 3 [ ] Paradigmas de aprendizagem com supervisão exigem maior intervenção humana do que qualquer outro paradigma uma vez que necessitam de quem desempenhe o papel de supervisor.

Verdadeiro (V)

- **Justificativa:** Nos paradigmas de aprendizagem supervisionada, é necessário ter dados rotulados, o que exige intervenção humana para classificar e organizar os dados previamente.

QUESTÃO 4 [v] O tratamento de valores nulos (*missing values*) existentes num *dataset* pode envolver a remoção de observações/registos ou de atributos/características.

Verdadeiro (V)

- **Justificativa:** Tratamento de valores nulos pode incluir tanto a remoção de registros que contenham valores ausentes quanto a exclusão de atributos que possuem muitos dados faltantes, dependendo da relevância para o modelo.

QUESTÃO 5 [f] A matriz de confusão à direita apresenta um valor de *accuracy* de $^{165}/_{150}$.

n=165

		PREVISÃO NÃO	PREVISÃO SIM	
ATUAL	NÃO	50	10	60
	SIM	5	100	105
		55	110	

Falso (F)

- **Justificativa:** A matriz de confusão não indica que o valor de *accuracy* seja $\frac{165}{150}$. Para calcular a *accuracy*, usa-se a fórmula:

$$\text{Accuracy} = \frac{\text{predições corretas}}{\text{total de predições}} = \frac{50 + 100}{165} = 0,91\ (91\%).$$

QUESTÃO 6 [f] Num processo de aprendizagem automática, a qualidade dos dados não afeta os resultados do processo uma vez que na fase de preparação de dados serão resolvidos todos os problemas como, por exemplo, ruído, *outliers*, dados falsos ou dados duplicados.

Falso (F)

- **Justificativa:** Embora a fase de preparação de dados seja fundamental, ela não pode resolver completamente problemas como ruído, outliers ou dados falsos, especialmente quando eles são persistentes ou complexos. A qualidade dos dados impacta diretamente o desempenho do modelo.