|  |  |
| --- | --- |
| Nome: Leandro de Jesus Luna | RM: 86492 |

Inteligência Artificial e Computacional

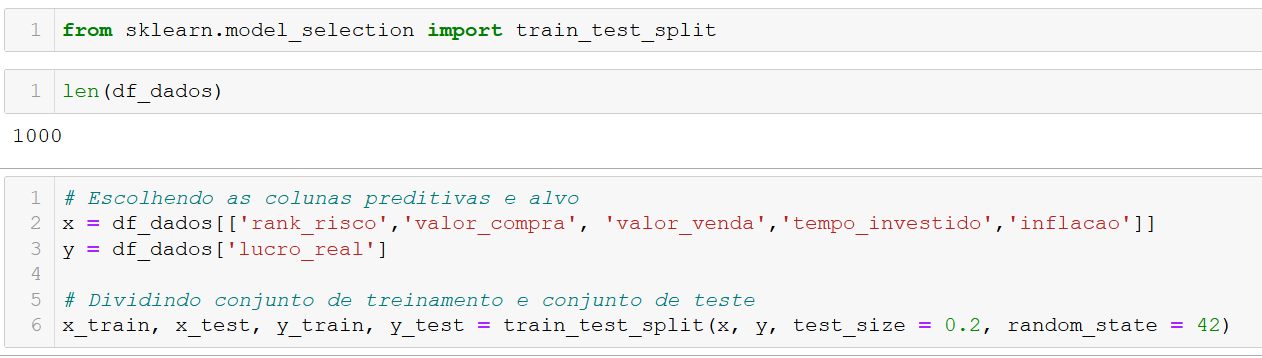
Checkpoint 6

## Instruções

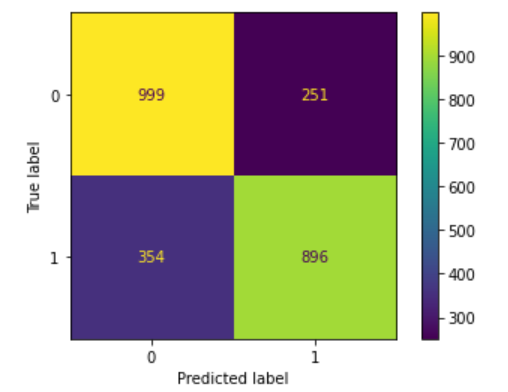
* Preencha esse questionário com suas respostas, sendo o mais claro possível e usando os termos técnicos aprendidos nas aulas;
* São **20 perguntas de múltipla escolha**, cada uma valendo 0.3 pontos, e **4 perguntas dissertativas** valendo 1 ponto cada;
* A atividade é **individual**;
* Envie como solução um arquivo .pdf devidamente preenchido;
* O nome do arquivo enviado deve ser **NAC5-NomeSobrenome.pdf**, onde **NomeSobrenome** deve ser trocado pelo nome e sobrenome da pessoa que está submetendo a atividade no Teams;

## Questões de Múltipla Escolha

Sobre o código abaixo, responda as questões 1, 2 e 3:



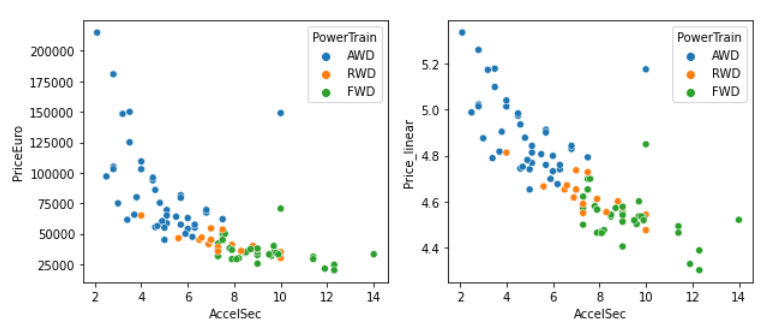
1. (0.3 ponto) O número de linhas na variável **x\_train** é:
2. 200
3. 500
4. 800
5. 1000
6. (0.3 ponto) O número de colunas descritivas utilizadas para treinamento é:
7. 1
8. 4
9. 3
10. 5
11. (0.3 ponto) É **correto** afirmar que:
12. O código foi desenvolvido em Python, no framework Jupyter Notebook, usando a biblioteca Scikitlearn.
13. O código foi desenvolvido em Java, no framework NetBeans, usando biblioteca Scikitlearn.
14. Esse código foi usado para se criar um chatbot sobre investimentos.
15. A função train\_test\_split é uma função nativa do NodeRED para separar as mensagens de teste da integração.
16. (0.3 ponto) Se temos dados descritivos e alvo, sendo nosso alvo uma string com três valores possíveis, é mais indicado começar utilizando a técnica de:
17. Aprendizado de máquina supervisionado de regressão.
18. Aprendizado de máquina supervisionado de classificação.
19. Aprendizado de máquina não supervisionado de agrupamento.
20. Aprendizado de máquina não supervisionado de redução de dimensionalidade.
21. (0.3 ponto) A imagem a seguir representa:



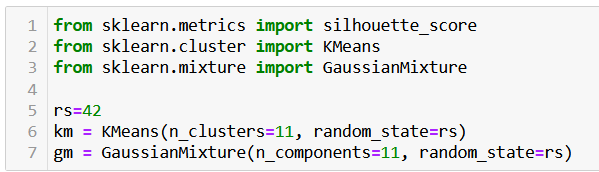
1. O resultado de um algoritmo de regressão;
2. O resultado da análise de desempenho de uma técnica não supervisionada;
3. A matriz de confusão de um regressor;
4. A matriz de confusão durante a análise de desempenho de uma classificação.
5. (0.3 pontos) Ao trabalharmos com Aprendizado de Máquina Supervisionado, **não devemos**:
6. Separar os dados em conjuntos de treinamento e teste.
7. Observar os tipos de dados e as estatísticas básicas dos dados antes de passá-los para um modelo de IA.
8. Passar como dados de treinamento as colunas descritivas e alvo.
9. Usar métricas adequadas para cada tipo de problema.
10. (0.3 ponto) Para criar um classificador de SPAM foram selecionados alguns algoritmos de aprendizado de máquina. Qual das opções abaixo **não é** um exemplo adequado de algoritmos para detecção de SPAM?
11. Árvore de Decisão e RandomForest;
12. K-vizinhos mais próximos (KNN) e Regressão Logística;
13. Regressão Linear e DBSCAN;
14. Máquina de Vetor Suporte (SVM) e K-vizinhos mais próximos (KNN).
15. (0.3 ponto) O comando abaixo calcula:



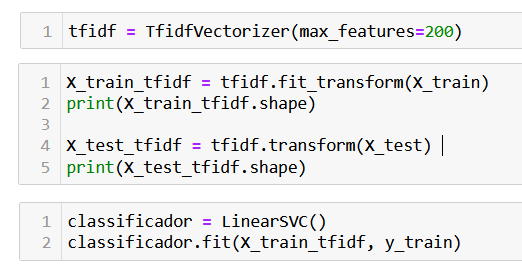
1. A divisão do dataframe dados pela coluna média (mean);
2. O agrupamento do dataframe CLASS\_LABEL;
3. A classificação do dataframe dados pela média do agrupamento da coluna CLASS\_LABEL;
4. A média de todos as colunas do dataframe dados agrupada pela coluna CLASS\_LABEL;
5. (0.3 ponto) Dado um antivírus baseado em Machine Learning, qual afirmação é **falsa**?
6. O algoritmo deve ser retreinado periodicamente adicionado novos exemplos de vírus e softwares legítimos;
7. O algoritmo pode ser do tipo supervisionado, funcionando como um classificador;
8. As features escolhidas para os códigos fontes representam fingerprints (impressões digitais) dos softwares;
9. São algoritmos estáticos, que se baseiam em um banco de dados prévio de todos os vírus conhecidos e as principais rotinas de ameaça, sem muita capacidade de generalização baseada em aprendizado estatístico.
10. (0.3 ponto) Os gráficos abaixo são:



1. Dois histogramas dos dados de preço em euro e linearizados em função da variável de aceleração (AccelSec), agrupados por tipo de tração (PowerTrain);
2. Dois boxplots separados pelo tipo de tração (PowerTrain) de carros;
3. Uma forma de representar dados linearizados e não linearizados em função do número de tração (AccelSec) do carro;
4. Duas dispersões agrupadas pelo tipo de tração (PowerTrain), sendo uma a função do preço pela aceleração (AccelSec) e a outra, do preço linearizado em função da aceleração.
5. (0.3 ponto) O termo maldição da dimensionalidade se refere a:
6. Lentidão com que os algoritmos supervisionados trabalham quando a quantidade de linhas no dataset aumenta drasticamente;
7. O fato de que é impossível para seres humanos enxergarem dados com mais de três colunas;
8. O fato de que, com o aumento do número de colunas, o espaço de definição do problema cresce exponencialmente e a amostragem disponível de exemplos do problema pode não ser suficiente para o aprendizado estatístico;
9. É um problema de overfitting devido a grande quantidade de exemplos disponíveis no dataset.
10. (0.3 ponto) Sobre as técnicas de representação de texto no contexto do processamento de linguagem natural (NLP), é **falso** afirmar que?
11. Bag of Words é uma técnica de sentence embedding que permite transformar frases em vetores numéricos;
12. BoW e W2V são usados para representar texto na forma de árvores semânticas que expressam as relações entre entidades através de grafos;
13. O W2V é uma técnica de word embedding que pode ser usada conjuntamente com operação vetores como produto interno e soma de vetores para a manipulação semântica;
14. O TF-IDF é uma técnica similar ao BoW, porém utilizando uma ponderação ranqueada para cada valor numérico que diminui a importância de palavras que aparecem repetidamente em dois datasets.
15. (0.3 ponto) Redes neurais artificiais são?
16. Técnicas clássicas de inteligência artificial desenvolvidas por Alan Turing e colaboradores;
17. Técnicas modernas de inteligência artificial desenvolvidas por Warren McCulloch e Walter Pitts no início dos anos 2000;
18. Técnicas de aprendizado de máquina baseadas na hipótese de Turing sobre inteligência artificial e neurônios biológicos;
19. Técnicas bioinspiradas de aprendizado de máquina cujo princípio de funcionamento baseia-se na descrição matemática de um neurônio;
20. (0.3 ponto) Em um sistema de controle baseado em Lógica Difusa podemos:
21. Criar regras precisas para lidar com valores semanticamente imprecisos mapeando tais valores para intervalos numéricos de pertinência;
22. Lidar com incerteza estatística expressando a aleatoriedade natural dos dados através de regras semânticas precisas;
23. Expressar regras imprecisas através de valores semânticos precisos que lidam com a aleatoriedade inerente dos dados de mundo real;
24. Capturar a precisão semântica da fala humana em regras bem definidas que mapeiam os valores sintáticos em intervalos de probabilidade.
25. (0.3 ponto) Sobre RNAs e IA Simbólica, é correto afirmar que:
26. RNAs são uteis para expressar conhecimento racional baseado em lógica formal;
27. Sistemas baseados em conhecimento podem possuir milhões de parâmetros devendo ser treinados por dias em GPUs;
28. Sistemas simbólicos apresentam a vantagem de escalabilidade e explicabilidade (explainability) e a desvantagem de não conseguirem extrapolar para dados fora do treinamento (OOD);
29. Sistemas neurais apresentam vantagem de aprendizado sob incerteza e escalabilidade e a desvantagem de não conseguirem extrapolar para dados fora do treinamento (OOD).
30. (0.3 ponto) O trecho de código a seguir é usado para:



1. Criar modelos de clusterização de dados;
2. Criar modelos de classificação de dados;
3. Criar modelos de predição de dados;
4. Criar modelos de análise exploratória dos dados.
5. (0.3 ponto) Durante o ano estudamos diversos tópicos sobre NLP, Chatbot e Inteligência Artificial. Sobre esses tópicos, é **incorreto** afirmar que:
6. Chatbots são ferramentas de interação humano-computador baseados em linguagem natural.
7. Node-RED é uma ferramenta de desenvolvimento baseada em JavaScript que serve para integrar serviços online (APIs) e tem muitas aplicações em Internet das Coisas (IoT).
8. Aprendizado de Máquina é um dos tipos de Inteligência Artificial no qual, os algoritmos utilizados não são codificados para resolver um problema específico, mas para extrair padrões de dados de maneira a aprender a solucionar o problema indiretamente, sem ter sido programado explicitamente para isso.
9. Os chatbots construídos usando Watson Assistant possuem intenções, entidades e variáveis de contexto. Intenções são palavras soltas, entidades são variáveis criadas para armazenar informação durante o diálogo e variáveis de contexto são frases longas que capturam a intenção do usuário que interage com o sistema.
10. (0.3 ponto) O código abaixo está:



1. Transformando dados numéricos em texto para treinar um classificador;
2. Criando um classificador de texto usando TF-IDF como embedding;
3. Criando um classificador de texto usando BoW como embedding;
4. Realizando o processamento de linguagem natural com um vocabulário de 200 frases para a transformação de dados de texto (strings) para números (float).
5. (0.3 ponto) É **correto** afirmar que:
6. Speech-to-Text é uma técnica de conversão de texto para fala usada para Sintetização de Fala.
7. Text-to-Speech é uma técnica de conversão de fala para texto usada para Reconhecimento de Fala.
8. Chatbots habilitados por voz devem necessariamente usar uma técnica end-to-end, na qual o áudio de entrada é processado diretamente, gerando um áudio de saída.
9. Reconhecimento de Fala é a técnica de transformar áudio contendo fala em um texto escrito, como em um ditado, enquanto que Reconhecimento de Voz é a técnica de identificar quem está falando em um diálogo (rotular os interlocutores).
10. (0.3 ponto) Sobre DeepLearning é **falso** afirmar que:
11. CNNs são redes neurais convolucionais usadas para majoritariamente para a classificação de áudio;
12. RNNs são redes recorrentes majoritariamente usadas para processamento de áudios e séries temporais;
13. GANs são redes adversárias generativas que podem ser usadas para criação de imagens falsas;
14. Redes Multilayer Perceptron são redes totalmente conectadas que podem ser usadas como camadas de classificação em CNNs.

Preencha as alternativas escolhidas das questões anteriores:

|  |  |
| --- | --- |
| Questão | Alternativa |
| 1 | C |
| 2 | D |
| 3 | A |
| 4 | B |
| 5 | D |
| 6 | C |
| 7 | C |
| 8 | D |
| 9 | D |
| 10 | D |
| 11 | C |
| 12 | B |
| 13 | D |
| 14 | A |
| 15 | D |
| 16 | A |
| 17 | D |
| 18 | B |
| 19 | D |
| 20 | A |

## Questões Dissertativas

1. (1 ponto) Quais as diferenças entre Inteligência Artificial e Aprendizado de Máquina?

R: Inteligência Artificial busca emular o pensamento humano, enquanto o Aprendizado de Máquina são algoritmos e tecnologias que permitem aos sistemas identificarem padrões, tomar decisões e se auto aperfeiçoarem, através dos dados.

1. (1 ponto) Explique o que é distribuição dos dados e sua relação com underfitting e overfitting. Como lidamos com esses problemas?

R: Distribuição dos dados são todos os possíveis valores - e intervalos – dos dados em um conjunto de dados (dataset). Através de regras de distribuição, pode-se inferir, por exemplo, a probabilidade de um dado ser da categoria X ou Y, bem como calcular sua frequência em um conjunto de dados.

Em relação ao underfitting e o overfitting, é justamente lidando com distribuição de dados que eles ocorrem. Se tivermos uma distribuição de dados muito desproporcional entre teste e treinamento, pode-se gerar um algoritmo de pouca generalização, ou muita generalização; e busca-se o meio termo desse cenário.

1. (1 ponto) Explique o que é um modelo pré-treinado, qual a relação dele com a RAM e o HD/SSD de um computador e de que forma podemos utilizar tais modelos.

R: Modelos pré-treinados são modelos previamente ‘capacitados’ – ou, propriamente dito, treinados – em um uma tarefa ou conjunto de dados. Tais modelos contribuem com um menor custo de carregamento de dados, evitando tanto gargalos de falta de memória RAM, quanto limitações de armazenagem em HD/SSD.

Desta maneira, é possível transportar um modelo previamente treinando de um host a outro, sendo muito utilizado em tarefas de CNN (Classificação de Imagens).

1. (1 ponto) Você recebeu a seguinte tabela de dados. Quais gráficos você faria e por quê? Qual técnica de aprendizado de máquina você usaria nesses dados e com qual objetivo? Que tipo de preocupações éticas e legais você deve ter ao trabalhar com esses dados?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Índice** | **Nome** | **Idade** | **Anos de Escolaridade** | **Remuneração Atual** | **Sexo** | **Estado Civil** | **Limite de Crédito** | |
| 0 | João | 27 | 13 | R$ 2200,00 | F | Sol. | R$ 500,00 | |
| 1 | Maria | 37 | 12 | R$ 4200,00 | F | Cas. | R$ 2000,00 | |
| 2 | Beatriz | 26 | 16 | R$ 2200,00 | M | Sol. | R$ 1000,00 | |
| 3 | Ronaldo | 29 | 4 | R$ 5000,00 | M | Div. | R$ 3000,00 | |
| ... | | | | | | | |
| 10000 | Vivian | 40 | 20 | R$ 20000,00 | F | Sol. | R$ 25000,00 | |

R:

* ‘’Quais gráficos você faria e por quê?’’ – Dentre diversas possibilidades, poderia ser gerado um histograma para visualizar a distribuição das diferentes colunas/features, bem como um gráfico de correlação de uma coluna/feature e outra. Através dessa análise exploratória inicial, pode-se pensar em diferentes abordagens no momento da elaboração de um algoritmo, nesse caso, supervisionado.
* ‘’Qual técnica de aprendizado de máquina você usaria nesses dados e com qual objetivo?’’ - Poderia ser utilizado uma técnica de regressão linear para estipular o valor do ‘limite de crédito’ – coluna preditiva - com base nas colunas descritivas: ‘Idade’, ‘Anos de Escolaridade’ e ‘Remuneração Atual’.
* ‘’Que tipo de preocupações éticas e legais você deve ter ao trabalhar com esses dados?’’ – Em relação a questões éticas, no conjunto de dados proposto, pode-se pensar em quais as profissões dos indivíduos que foram expostos nesse conjunto. Em um cenário onde, um individuo do sexo feminino tem um limite de crédito maior do que um individuo do sexo masculino, não necessariamente expõe que os indivíduos do sexo masculino têm uma profissão desvalorizada em relação aos indivíduos do sexo feminino. Deve-se, também, levar em conta o espectro social desses indivíduos para chegar em uma conclusão definitiva.  
    
  Em relação a questões legais, trata principalmente da lei de sigilosidade dos dados; em outros termos, a privacidade de divulgação e exposição dos dados (LGPD).