Universidade Federal do Pará
Instituto de Ciências Exatas e Naturais
Faculdade de Computação
Disciplina: Inteligência Artificial
Professor: Jefferson Morais
Monitor: Helder Matos

Instruções para a entrega do trabalho

Neste trabalho, você deve criar e entregar um um único documento no formato **Jupyter Notebook** (.ipynb) utilizando a linguagem de programação Python. O Jupyter Notebook é uma aplicação web que permite criar e compartilhar documentos com código-fonte, visualizações e explicações em texto. O notebook deve conter células de código que implementam as tarefas a seguir e células de explicação em markdown. Você pode criar um Jupyter Notebook localmente, instalando o Jupyter via Anaconda ou pip, ou na nuvem, utilizando plataformas como **Google Colab**. O Colab permite armazenamento automático no Google Drive e facilita o compartilhamento.

O trabalho pode ser realizado **individualmente ou em grupos de até quatro pessoas**, com **pontuação máxima de 8 pontos**, conforme a qualidade da entrega. O prazo de entrega é **23 de outubro de 2024 às 23:59h**.

Para a entrega, um único membro do grupo deve enviar um único arquivo pelo SIGAA, listando os nomes dos integrantes da equipe. Se usar plataformas como o Colab, gere um link de compartilhamento público, cole em um arquivo de texto e envie o arquivo. Se criar localmente, compacte o arquivo .ipynb em um .zip para envio. Siga essas orientações para garantir uma entrega organizada.

Garanta que a base de dados descrita na tarefa seja carregada automaticamente no seu notebook, anexando a mesma no arquivo zipado ou realizando o download do arquivo no Colab. Lembre-se que este é o trabalho final da disciplina, e é importante que você demonstre atenção à qualidade e ao zelo na execução.

Trabalho prático: teste vocacional na área de dados

O State of Data Brazil 2022 é um estudo abrangente sobre o mercado de trabalho brasileiro na área de dados, conduzido pela comunidade Data Hackers e a consultoria Bain & Company. A pesquisa foi realizada entre outubro e novembro de 2022, com 4271 respondentes, mapeando perfis demográficos, formação, remuneração, desafios profissionais e o impacto do trabalho remoto.

O dataset original era composto por 4271 instâncias e 353 atributos. Após seleção e renomeação de alguns atributos, o dataset foi reduzido para 14 atributos listados na tabela abaixo:

#	Nome do atributo	Descrição	Tipo
1	idade	Idade do respondente	Inteiro
2	genero	Gênero do respondente	String
3	etnia	Etnia do respondente	String
4	pcd	O respondente é Pessoa com Deficiência?	String
5	vive_no_brasil	O respondente mora no Brasil?	Booleano
6	estado_moradia	Se mora no Brasil, qual o estado?	String
7	nivel_ensino	Nível de instrução do respondente	String
8	formacao	Formação acadêmica do respondente	String
9	tempo_experiencia_dados	Tempo de experiência do respondente na área de dados	String
10	linguagens_preferidas	Linguagens de programação preferidas do respondente	String
11	bancos_de_dados_preferidos	Ferramentas e sabores de bancos de dados preferidos do respondente	String
12	cloud_preferida	Ferramentas de cloud computing preferidas do respondente	String
13	cargo	Cargo atual ocupado pelo respondente	String

Você pode fazer o download do dataset neste link: https://bit.ly/facomp-sods.

Utilize a base de dados fornecida e aplique o processo KDD para realizar uma tarefa de **classificação de dados**, sendo livre o uso de bibliotecas e frameworks. A classificação consiste no uso dos 12 primeiros atributos (de *idade* até *cloud_preferida*) para encontrar qual o *cargo* ideal para um candidato a uma vaga na área de dados. As tarefas de cada etapa do processo KDD a serem cumpridas estão descritas a seguir.

Seleção

Escolha as instâncias e os atributos mais relevantes para a análise, justificando claramente o motivo de cada escolha. Se optar por utilizar todas as 4271 linhas e 13 atributos disponíveis, essa decisão deve ser explicitada com as devidas justificativas.

Realize uma análise exploratória dos dados utilizando tabelas, gráficos ou outras ferramentas de visualização adequadas. Essas ferramentas devem ajudar a compreender melhor os atributos, suas distribuições e a identificar possíveis erros ou inconsistências.

Pré-processamento

Remova ou trate dados faltantes, incorretos ou inconsistentes para melhorar a qualidade geral do conjunto de dados. Isso inclui eliminar ruídos que possam comprometer a análise.

Além disso, crie novos atributos com base nos já existentes, revelando informações úteis que possam estar ocultas (ex.: transformar o tempo de experiência em um valor inteiro pode facilitar a modelagem).

Formatação

Converta os dados originais para um formato que seja compatível com as técnicas de modelagem a serem utilizadas (consulte a subseção de Mineração de Dados deste documento). Certas técnicas, como redes neurais artificiais, exigem a transformação de atributos categóricos em valores numéricos.

Implemente a técnica de **Hold-out estratificado 70%-30%** (mantendo a proporção de classes), dividindo o conjunto de dados em subconjuntos de treinamento e teste.

Mineração de dados

Escolha e justifique o uso de **duas e somente duas** das seguintes técnicas:

- Regressão Logística (Logistic Regression)
- Árvore de Decisão (Decision Tree)
- Florestas Aleatórias (Random Forests)
- Máquina de Vetor de Suporte (Support Vector Machine)
- Redes Neurais Artificiais (Artificial Neural Networks)

Justifique também a escolha dos hiperparâmetros utilizados, explicando por que esses parâmetros foram considerados os mais adequados para treinar os modelos.

Ajuste os modelos escolhidos utilizando os dados de treinamento e, em seguida, extraia as respectivas matrizes de confusão.

Avaliação

Utilizando o conjunto de treinamento, calcule a **acurácia geral** de ambos os modelos. Com base nesses resultados, analise, destaque e compare os modelos computacionais. Use uma abordagem criativa para comparar os classificadores, evidenciando qual algoritmo demonstrou melhor desempenho nos experimentos.

Indique qual técnica seria mais adequada para classificar o conjunto de dados escolhido, justificando suas conclusões com base em evidências concretas.