

 UNIVERSIDADE FEDERAL DE SANTA CATARINA		Campus Araranguá Rodovia Governador Jorge Lacerda, 3201, Jardim das Avenidas Araranguá - Santa Catarina – Brasil / CEP 88900-000 www.ararangua.ufsc.br / +55 (48) 3721.6448		1/3
Avaliação II				
Disciplina	Tópicos Especiais III (DEC7553)		08655	2022.2
Aluno(a)s				

1) Relacione os conceitos da 1ª coluna com as respectivas definições na 2ª coluna – (1,5):

[1]	Descoberta de Conhecimento em Bases de Dados	[]	Cria árvores de decisão a partir da seleção randômica de exemplos dos dados e obtém a predição de cada árvore selecionando a melhor solução.
[2]	Pré-processamento	[]	Conceito criado para unificar a estatística, análise de dados, aprendizado de máquina e seus métodos relacionados visando entender e analisar fenômenos reais com dados.
[3]	Aprendizado de Máquina	[]	Processo não trivial, interativo e iterativo, visando a identificação de padrões compreensíveis, que sejam válidos, novos, potencialmente úteis a partir de grandes conjuntos de dados.
[4]	Aprendizado Supervisionado	[]	Objetiva descobrir um relacionamento entre os atributos precursores e o atributo meta, usando registros cuja classe é conhecida, para se construir um modelo que possa ser aplicado a objetos ainda não classificados.
[5]	Aprendizado Não Supervisionado	[]	Processo pela qual determinado conjunto de dados de teste é dividido em N subconjuntos e a partir desses N subconjuntos são realizados N testes.
[6]	Avaliação	[]	Expressa a distância, similaridade, correlação ou associação entre dois objetos quaisquer.
[7]	Pós-processamento	[]	Proporção de exemplos de uma classe (neste caso valores positivos) que foram corretamente classificados pelos exemplos classificados incorretamente na classe - $TP / (TP + FP)$.
[8]	Tarefa de Classificação	[]	Estrutura que, na tarefa de classificação envolvendo classes de valores discretos, permite estimar a qualidade dos resultados obtidos.
[9]	Tarefa de Agrupamento	[]	Processo pelo qual determinado algoritmo de aprendizado (indutor) recebe um conjunto de exemplos de treinamento para os quais os rótulos da classe associada são conhecidos e, a partir disso, constrói um modelo visando a correta determinação da classe para novos exemplos ainda não rotulados.
[10]	Tarefa de Associação	[]	Consiste na localização de materiais (usualmente documentos) de natureza não estruturada (usualmente textos) que satisfazem determinada necessidade por informação a partir de uma grande coleção.
[11]	Árvore de Decisão	[]	Técnica baseada na noção intuitiva de que determinados padrões de interação, representados na forma de um grafo, são importantes características capazes de descrever o comportamento de unidades de análise.
[12]	ID3	[]	Objetiva descobrir relacionamentos importantes em um conjunto de dados, tal que, a presença de um item em uma determinada transação irá implicar na presença de outro item na mesma transação.
[13]	Cross-validation	[]	Processo pelo qual o algoritmo indutor analisa os exemplos fornecidos e tenta determinar quais destes podem ser agrupados de alguma maneira, formando agrupamentos ou <i>clusters</i> .

 UNIVERSIDADE FEDERAL DE SANTA CATARINA		Campus Araranguá Rodovia Governador Jorge Lacerda, 3201, Jardim das Avenidas Araranguá - Santa Catarina – Brasil / CEP 88900-000 www.ararangua.ufsc.br / +55 (48) 3721.6448		2/3
Avaliação II				
Disciplina	Tópicos Especiais III (DEC7553)		08655	2022.2
Aluno(a)s				

[14]	Matriz de Confusão	[]	Medida que avalia a frequência com que determinado nodo aparece no caminho mais curto entre dois nodos quaisquer.
[15]	Precisão (<i>Precision</i>)	[]	Etapa que visa a preparação e a transformação de dados de modo que possam ser utilizados por algoritmos de mineração de dados e, deste modo, conduzir à descoberta de conhecimento.
[16]	Medidas de Similaridade	[]	Técnica de classificação que, através de um processo de otimização heurística e baseado na teoria da informação, seleciona iterativamente determinadas variáveis que promovem a melhor separação de classes de acordo com alguma função de custo, visando criar regiões disjuntas e estabelecendo uma fronteira de decisão.
[17]	Centralidade de Proximidade	[]	Processo pelo qual se objetiva aferir a qualidade da aplicação/utilização de algoritmos de Mineração de Dados sobre determinado conjunto de dados.
[18]	<i>Random Forest</i>	[]	Objetiva servir como um passo anterior à tarefa de classificação quando não se possui um conjunto de dados previamente classificado, possibilitando a reunião de itens semelhantes em determinado grupo.
[19]	Ciência de Dados	[]	Um programa aprende a partir da experiência E, em relação a uma classe de tarefas T, com medida de desempenho P, se seu desempenho em T, medido por P, melhora com E.
[20]	Dendrograma	[]	Etapa do processo de KDD que consiste na avaliação dos resultados obtidos, isto é, analisa se o conhecimento descoberto é relevante ou não.
[21]	Grafo	[]	Determina a porcentagem de determinado <i>itemset</i> dentre todas as transações da Base de Dados.
[22]	Análise de Rede Social	[]	Representa a distância natural entre todos os pares de nodos, definida pelo tamanho dos caminhos mais curtos.
[23]	Centralidade de Intermediação	[]	Árvore que iterativamente divide o conjunto de dados em subconjuntos menores até que cada subconjunto consista de somente um objeto.
[24]	Recuperação de Informação	[]	Algoritmo baseado na teoria da informação que constrói uma árvore de decisão onde cada vértice (nodo) corresponde a um atributo, e cada aresta da árvore a um valor possível do atributo.
[25]	Medida do Suporte	[]	Conjunto de vértices conectados por arestas que podem ser direcionadas ou não.

 UNIVERSIDADE FEDERAL DE SANTA CATARINA		Campus Araranguá Rodovia Governador Jorge Lacerda, 3201, Jardim das Avenidas Araranguá - Santa Catarina – Brasil / CEP 88900-000 www.ararangua.ufsc.br / +55 (48) 3721.6448		3/3
Avaliação II				
Disciplina	Tópicos Especiais III (DEC7553)		08655	2022.2
Aluno(a)s				

- 2) Considerado o conjunto `tae.csv` e, utilizando a linguagem Python e as bibliotecas apresentadas na disciplina, elabore uma árvore de decisão. Após isso reduza a profundidade da árvore. Apresente a acurácia inicial e após a redução da profundidade. Também apresente duas regras geradas pela árvore. O conjunto de dados representa avaliações de desempenho no ensino ao longo de alguns semestres e possui as colunas 'ta_native', 'course_instr', 'course', 'summer_regular', 'class_size' e 'label'. A coluna 'label' representa o atributo meta, ou seja, o objetivo da classificação - (1,5).
- 3) Utilizando o conjunto de dados (`wine.csv`) e, utilizando a linguagem Python e as bibliotecas apresentadas na disciplina, elabore um algoritmo de aprendizado de máquina do tipo *Random Forest*. Na sequência calcule a contribuição de cada característica e realize novamente o treinamento e o teste sem algumas das características menos importantes. Considerando este novo conjunto de dados elabore o algoritmo *k-means* aplicando uma transformação nos dados. Ao final apresente as acurácias dos dois algoritmos, *Random Forest* e *K-means*. O conjunto de dados representa análises de vinho e possui as colunas 'label', 'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash', 'Magnesium', 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins', 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'. A coluna 'label' representa o atributo meta, ou seja, o objetivo da classificação. A coluna 'Proline' deve ser descartada para todas as análises – (2,0).
- 4) Utilizando o conjunto de dados '*breast cancer*' através do método `load_breast_cancer()` disponível na biblioteca `sklearn.datasets`, realize uma análise utilizando algum dos algoritmos de classificação ou agrupamento estudados. Ao final apresente a acurácia – (1,5).
- 5) Elabore um grafo com um conjunto de nodos (mais de 15) em que fique clara a separação em três grupos. Os grupos devem estar conectados somente por um dos nós de cada grupo. Após isso calcule as métricas de centralidade de intermediação (*betweenness centrality*) e centralidade de proximidade (*closeness centrality*). Ao final apresente uma relação dos nós (5 nós, por exemplo) mais importantes ordenando do mais relevante para o menos relevante considerando as métricas de centralidade de intermediação e proximidade - (1,5).
- 6) Elabore um código em Python para um sistema simples de recuperação e recomendação de informações respeitando os seguintes passos – (2,0):
 - a) Elabore uma matriz Documento X Termo (por exemplo, 10 documentos e 15 termos) com pesos já normalizados entre 0.0 e 1.0. Distribua os pesos de maneira esparsa;
 - b) A partir da matriz calcule uma matriz de similaridades entre os documentos utilizando a métrica do cosseno;
 - c) Por fim, para cada documento identifique os $k=3$ documentos mais similares (desconsidere o próprio documento) apresentando o identificador do documento e a similaridade calculada. Para cada documento, os $k=3$ documentos mais relacionados devem estar ordenados de maneira decrescente, ou seja, da maior similaridade para a menor.