**APRENDIZAGEM DE MÁQUINA (MAQ-04)**

**AVALIAÇÃO P3 – 04/07/2025**

**1) ANÁLISE DE DADOS (1,0 ponto)**

Considerando o conjunto de dados contendo informações sobre fogo em florestas (forestfires.csv em anexo): e calcule as seguintes estatísticas sobre os dados de cada atributo:

1. Calcule, para cada coluna desse conjunto de dados, as estatísticas univariadas pertinentes dentre frequências, moda, média, mediana, desvio-padrão, Q1, Q3, obliquidade e curtose.
2. Responda: qual é a variável com maior dispersão em torno de sua média?
3. Calcular a matriz de correção para os atributos quantitativos presentes.
4. Responda: qual dupla de atributos quantitativos tem a maior dependência em si?

**2) AGRUPAMENTO DE CLASSES (2 pontos)**

Considere o código abaixo referente a um classificador que utiliza o método de agrupamento para estimar os níveis de obesidade de pessoas com base em hábitos alimentares e condições físicas. Responda:

1. Quantos atributos são disponibilizados para cada amostra do conjunto de dados?
2. Quais são as categorias de classificação?
3. Por que é necessário utilizar o método fit\_transform da classe LabelEncoder para alguns dos atributos do conjunto de dados?
4. Qual a acurácia do classificador para os dados de treinamento?
5. Qual é classe predita pelo algoritmo para uma pessoa com as seguintes características:
   1. Sexo feminino (Gender)
   2. 42 anos (Age)
   3. 1.75 cm de altura (Height)
   4. 75 kg de peso (Weight)
   5. Sem histórico familiar de sobrepeso (family\_history\_with\_overweight)
   6. Que se alimenta de comida com alta calorias frequentemente (FAVC)
   7. Que não ingere vegetais nas refeições (FCVC)
   8. Que faz 3 refeições principais no dia (NCP)
   9. Que as vezes como algo entre as refeições (CAEC)
   10. Que não fuma (SMOKE)
   11. Que bebe 2 litros de água por dia (CH20)
   12. Que não monitora o número de calorias ingeridas no dia (SCC)
   13. Que nunca faz atividades física (FAF=0)
   14. Que nunca usa dispositivos digitais (TUE=0)
   15. Que as vezes ingere bebida alcoólica (CALC)
   16. Que usa transporte público (MTRANS)
6. Caso seja definido um hiperparâmetro do algoritmo para que sejam utilizados 2 vizinhos na predição, a acurácia dos dados de treinamento irá melhorar ou não? Faça modificações no código e explique o resultado obtido comparando as acurácias obtidas.

from ucimlrepo import fetch\_ucirepo

from sklearn.preprocessing import LabelEncoder

from sklearn.neighbors import KNeighborsClassifier

# fetch dataset

estimation\_of\_obesity\_levels\_based\_on\_eating\_habits\_and\_physical\_condition = fetch\_ucirepo(id=544)

# data (as pandas dataframes)

X = estimation\_of\_obesity\_levels\_based\_on\_eating\_habits\_and\_physical\_condition.data.features

y = estimation\_of\_obesity\_levels\_based\_on\_eating\_habits\_and\_physical\_condition.data.targets

# metadata

print(estimation\_of\_obesity\_levels\_based\_on\_eating\_habits\_and\_physical\_condition.metadata)

# variable information

print(estimation\_of\_obesity\_levels\_based\_on\_eating\_habits\_and\_physical\_condition.variables)

le = LabelEncoder()

col\_cat = ['Gender', 'family\_history\_with\_overweight', 'FAVC', 'CAEC', 'SMOKE', 'SCC', 'CALC', 'MTRANS']

for column\_name in col\_cat:

    X[column\_name] = le.fit\_transform(X[column\_name])

y = le.fit\_transform(y)

knnClassifier = KNeighborsClassifier(metric = 'minkowski', p=2)

knnClassifier.fit(X, y)

**3) APRENDIZAGEM NÃO SUPERVISIONADA (1,0 ponto)**

Altere o código do exercício anterior para utilizar um algoritmo de aprendizagem não supervisionada estudado em aula, de maneira que os dados sejam agrupados em diferentes grupos. Explique quantos grupos você escolheu como hiperparâmetro e ao final imprima os valores dos centroides de cada grupo.

Configure o hiperparâmetro para definir a semente de geração de número pseudoaleatórios utilizada no treinamento do algoritmo, utilizando como número a soma das primeiras letras do seu nome e último sobrenome. Por exemplo, para José de Almeida, o número de semente será 10+1, ou seja, “J” = 10 e “A” = 1.

**4) CLASSIFICADOR BINÁRIO (2,5 pontos)**

Utilizar o algoritmo SGDClassifier para diferenciar tipos de cogumelos em classes de venenosos e não venenosos, com base em suas características.

Para tanto, o conjunto de dados criado pela *Audobon Society Field Guide*, disponibilizado no endereço <https://archive.ics.uci.edu/dataset/73/mushroom> , deverá ser utilizado.

Este conjunto de dados contém 8.124 amostras que deverão ser dividas adequadamente em subconjuntos para treinamento do algoritmo. Cada amostra apresenta 22 características de 23 espécies de cogumelos, que deverão ser tratados para permitir o seu processamento.

Dividir o conjunto de dados em subconjuntos de treinamento e teste, na proporção de 70% e 30%, respectivamente. Para tanto, utilizar o número **51** para a semente de geração de número pseudoaleatórios.

Ao final do treinamento, computar o F1 score, disponibilizando-o juntamente com o código fonte desenvolvido.

**5) PREDIÇÃO COM REDES NEURAIS ARTIFICIAIS (3,5 pontos)**

Treinar uma Rede Neural Artificial (RNA) do tipo Perceptron Multicamadas para estimar o número de ocupantes de um quarto, com base na leitura 7 sensores (tipos de sensores: temperatura, luz, som do ambiente, nível de CO2 e sensor passivo infravermelho).

Para tanto, o conjunto de dados disponibilizado no endereço <https://archive.ics.uci.edu/dataset/864/room+occupancy+estimation> deverá ser utilizado, que consiste em 10.129 amostras coletadas a cada 30 segundos.

Este conjunto de dados será divido adequadamente em subconjuntos para treinamento do algoritmo.

A arquitetura da RNA deverá conter pelo menos uma camada oculta.

Juntamente com o código fonte desenvolvido, disponibilizar:

1. As curvas de acurácia e da função custo ao longo das épocas de aprendizado, informando, se for o caso, em que época do treinamento começou a ocorrer sobreajustamento (*overfitting*).
2. A matriz de confusão com base nos dados de teste.