

### Lista #3

**Curso: Ciência da Computação**

**Disciplina: Inteligência Artificial**

**Prof<sup>a</sup>. Cristiane Neri Nobre**

**Data de entrega: 09/03**

**Valor: 1 ponto**

#### Questão 01

---

Considerando o exercício da lista anterior sobre a base do TITANIC, gere a árvore de decisão com o hiperparâmetro **criterion = Entropy** e **criterion=Gini**

Explique como é o funcionamento do critério Gini e mostre os cálculos são obtidos. Compare as árvores obtidas pelos dois critérios.

#### Questão 02

---

Considerando ainda a mesma base de dados do TITANIC, investigue o impacto de outros hiperparâmetros da árvore tais como: **max\_depth**, **max\_features**, **min\_samples\_leaf**, dentre outros.

Discuta cada hiperparâmetro.

#### Questão 03

---

Investigue o funcionamento dos seguintes otimizadores de hiperparâmetros dos algoritmos de aprendizado de máquina:

**GridSearchCV** (from sklearn.model\_selection import **GridSearchCV**)

**RandomizedSearchCV** (rom sklearn.model\_selection import **RandomizedSearchCV**)

**BayesSearchCV** (from skopt import **BayesSearchCV**)

E investigue, implemente e discuta o funcionamento deles para otimização dos hiperparâmetros para a base do TITANIC.

#### Questão 04

---

Considere um modelo de classificação binária que identifica fraudes em transações financeiras. Suponha que a base de dados tenha um número significativamente maior de transações legítimas do que fraudulentas. Com base nas métricas de avaliação **precisão (precision)**, **revocação (recall)** e **F1-score**, analise as seguintes afirmações:

I. Se o modelo tem alta **precisão**, isso significa que a maioria das transações classificadas como fraudulentas realmente são fraudes, mas pode estar deixando muitas fraudes reais passarem despercebidas.

II. Se o modelo tem alta **revocação**, isso significa que ele consegue identificar quase todas as fraudes, mas pode incluir muitas transações legítimas como fraudulentas.

III. O **F1-score** é útil quando há um grande desequilíbrio entre classes, pois equilibra precisão e revocação, sendo sempre a média aritmética dessas métricas.

Qual das alternativas abaixo é **correta**?

- A) Apenas I e II
- B) Apenas II e III
- C) Apenas I e III
- D) I, II e III

#### Questão 05

---

Um modelo de **diagnóstico de doenças raras** foi desenvolvido para identificar pacientes infectados com uma condição grave. Com base nas métricas **precisão (precision)** e **revocação (recall)**, analise as seguintes afirmações:

I. Se a **revocação** for aumentada, mais casos reais da doença serão detectados, mas isso pode aumentar os **falsos positivos**, reduzindo a precisão.

II. Se um modelo tem **alta precisão**, isso significa que a maioria dos pacientes diagnosticados como positivos realmente tem a doença, mas isso não garante que todos os doentes tenham sido identificados.

III. Para um diagnóstico de doenças **altamente letais**, um modelo com alta **precisão** sempre é preferível a um modelo com alta **revocação**, pois evita alarmes falsos e diagnósticos errados.

Qual das alternativas abaixo é **correta**?

- A) Apenas I e II
- B) Apenas I e III
- C) Apenas II e III
- D) I, II e III

#### Questão 06

---

O algoritmo de árvore de decisão C45 possui algumas diferenças em relação ao ID3. Quais são estas diferenças? Explique.

#### Questão 07

---

Para selecionar o melhor atributo na construção da árvore, o algoritmo de árvore de decisão C45 utiliza a **Razão de Ganho** (Gain Ratio) ao invés do ganho de informação utilizada pelo algoritmo ID3.

Explique as diferenças entre estas 2 medidas.