

|  |  |
| --- | --- |
| Colegio Universitario **IES** *Siglo 21* | |
| EXAMEN PARCIAL 1 | |
| **Materia: Introducción a la IA** | **Docente:** Ricardo Piña |
| **Modalidad: Presencial Semipresencial Distancia** | **Fecha:** |

Reservado para el alumno

|  |  |
| --- | --- |
| **Alumno**: Lautaro Santos Da Silveira | **Carrera:  INTELIGENCIA ARTIFICIAL Y CIENCIA DE DATOS** |
| **DNI**: 43 879 787 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Actividad** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **Total puntos** | NOTA |
| **Puntaje** |  |  |  |  |  |  |  |  |  |  |  |  |

* Objetivos:
  + Comprender los elementos fundamentales de la forma de analizar problemas de Aprendizaje.
  + Conocer las distintas clasificaciones de problemas de IA.
  + Ser capaz de aplicar a problemas cercanos a la realidad algún modelo de Machine Learning, seleccionar los hiperparámetros y evaluar el modelo.
* Modalidad de Evaluación:
  + Escrito e individual.
* Puntaje:
  + - El puntaje se determinará a través de la siguiente escala:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Puntaje | 1-24 | 25-39 | 40-54 | 55-61 | 62-66 | 67-72 | 73-79 | 80-87 | 88-95 | 96-100 |
| Nota | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Valoración | Logro  No satisfactorio | | | Logro básico | Logro satisfactorio | | Logro  Destacado | Logro  Sobresaliente | | Logro  Excelente |

PARA REMITIR EL EXAMEN TEÓRICO:

Guarde el examen teórico con *suApellido\_suNombre.docx* y remítalo por el servicio de Mensajería del Aula Virtual, antes de las 23 hs.

1. En los siguientes casos indique:

* **a qué tipo de Aprendizaje** (Supervisado, No Supervisado, etc) corresponde y
* **qué tipo de problema es** (Clasificación, Regresión, etc)(1 punto) :
  1. Se desea pronosticar el peso que tendrán las peras una vez maduras y se conocen los datos: alto, largo, cantidad\_días\_cosechada, peso\_madura para 1000 peras.
* Tipo de aprendizaje **SUPERVISADO** y es un problema de **REGRESIÓN**
  1. En un sistema de vehículo autónomo (es decir que se conduce por sí solo) es fundamental que el sistema pueda detectar si un ser humano se encuentra en la calle. **Se cuenta con miles de fotografías de personas y con miles de fotografías de elementos que suelen aparecer en la calle y que no son personas**. Se desea que nuestro sistema tome imágenes con una cámara y decida si lo que se ve en la imagen incluye o no a una persona.
* Tipo de aprendizaje **SUPERVISADO** y es un problema de **CLASIFICACIÓN**
  1. La gerencia de marketing de un hipermercado desea elaborar una campaña de marketing: desea enviar mensajes por mail o WhatsApp ofreciendo sólo los productos que los clientes podrían estar interesados en comprar. Para ello cuenta con la base de datos de las compras anteriores e información demográfica de cada uno de los clientes. La idea es que clientes *similares* comprarán productos *similares*. Nuestro objetivo es indicarle al gerente de mkt qué clientes se parecen más entre sí en función de los datos mencionados.
* Tipo de aprendizaje **NO SUPERVISADO** y es un problema de **CLUSTERING**

1. SITUACIÓN: (2 puntos)

Sea un problema de **Aprendizaje Supervisado** para resolver el cual, Ud posee un Conjunto de Datos C (que es sólo una fracción de todos los casos posibles o universo) y lo divide en dos (sub)conjuntos de datos A y B, de tal manera que C = A U B. Ud decide aplicar un modelo (como por ejemplo árboles de decisión).

En cada uno de los siguientes casos **comente** cada una de las alternativas y analice **Exactitud (Accuracy),** **Overfiting** que se obtendría en cada caso. Cuál de las estrategias utilizaría Ud?

* 1. Caso 1: Entrenar al modelo con C y testearlo sobre C

Nuestra **Accuracy** será de 1.00, es decir del 100%, lo que se resume a que tendremos **Overfiting**, ya que el programa solo tomará como verdadera los datos de C, y si ingresamos futuros datos nos dará error. A su vez está mal ya que no podemos entrenar con una variable y entrenar con la misma.

* 1. Caso 2: Entrenar al modelo con A y testearlo sobre C

En este caso, nuestro **Accuracy** tal vez nos de un valor razonable, pero tendríamos muy poco **Overfiting**, ya que entrenamos al programa con menos datos que con lo que lo testeamos. Esto se traduce a que el programa no responderá bien a los futuros pronósticos.

* 1. Caso 3: Entrenar al modelo con A y testearlo sobre B

En este caso nuestro **Accuracy** nos dará un valor razonable y podríamos usarlo, pero estaríamos entrenando al programa con B, por lo que nos daría una información errónea, y nuestro **Overfiting** dependería de la profundidad del modelo que decidamos usar.

* 1. Yo usaría la siguiente estrategia:

De los datos en A los subdividiría de nuevo en dos, validation test A y validation train A, teniendo mas datos el train. En los validation buscaría la profundidad adecuada comparando en diferentes modelos el que tenga el **Accuracy** mas alto. Una vez encontrado entrenaría con la totalidad de los datos de A con el árbol con la profundidad que encontré anteriormente, y testearia con los datos de B, teniendo A mas datos que B, asi encontraría el **Accuracy** real de mi programa. Una vez lo encuentro entrenaría con C, generando asi el programa que me pronostique con una gran cantidad de datos.

1. Dado el siguiente conjunto de datos donde **y** es la variable a pronosticar, **x1, x2, x3** son las características, features, o variables explicativas, aplique el método de **Árbol de Decisión con Ganancia de Información** para calcular manualmente el primer nodo mediante el método de la entropía. (**Nota**: si le es de utilidad, estos datos se encuentran en el archivo: Parcial\_1\_Ejs.xlsx) (2 puntos)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | x1 | x2 | x3 | y |
| 1 | No | Bajo | Positivo | **A** |
| 2 | Si | Alto | Negativo | **B** |
| 3 | Si | Alto | Positivo | **A** |
| 4 | Si | Medio | Negativo | **A** |
| 5 | Si | Alto | Negativo | **A** |
| 6 | No | Bajo | Negativo | **A** |
| 7 | No | Bajo | Negativo | **A** |
| 8 | No | Medio | Negativo | **B** |
| 9 | No | Bajo | Positivo | **B** |
| 10 | No | Bajo | Positivo | **A** |
| 11 | Si | Alto | Positivo | **A** |
| 12 | Si | Alto | Positivo | **B** |
| 13 | Si | Alto | Negativo | **B** |
| 14 | Si | Medio | Positivo | **A** |
| 15 | No | Medio | Negativo | **A** |

Escriba las cuentas aquí mismo hechas en Word, no importa si la simbología matemática no es precisa, por ejemplo si necesita escribir “logaritmo en base 2 de 40”, por ejemplo, puede escribirlo así: log2(40) que lo entenderemos!

Entropía inicial: cantidad de A= 10/15

Cantidad de B= 5/15

S= -[10/15 \* log2(10/15) + 5/15 \* log2(5/15)]

S= 0.91

**Entropía x1:**

**No=**

* Casos: 7
* Prop al nodo sup: 7/15
* Cant A: 5
* Cant B: 2

S de no= -[5/7 \* log2(5/7) + 2/7 \* log2(2/7)] = 0.86

**Si=**

* Casos: 8
* Prop al nodo sup: 8/15
* Cant A: 5
* Cant B: 3

S de si= -[5/8\* log2(5/8) + 3/8 \* log2(3/8)] = 0.95

**S X1= 7/15\*0.86+8/15\*0.95= 0.90**

**ENTROPIA DE X2:**

**Bajo=**

* Casos: 5
* Prop al nodo sup: 5/15
* Cant A: 4
* Cant B: 1

S de Bajo= -[4/5\* log2(4/5) + 1/5 \* log2(1/5)] = 0.72

**Medio=**

* Casos: 4
* Prop al nodo sup: 4/15
* Cant A: 3
* Cant B: 1

S de Medio= -[3/4\* log2(3/4) + 1/4 \* log2(1/4)] = 0.81

**Alto=**

* Casos: 6
* Prop al nodo sup: 6/15
* Cant A: 3
* Cant B: 3

S de Alto= -[3/6\* log2(3/6) + 3/6 \* log2(3/6)] = 1

**S X2= 6/15\*1+ 4/15\*0.81 + 5/15\*0.72= 0.85**

**ENTROPIA DE X3:**

**Positivo=**

* Casos: 7
* Prop al nodo sup: 7/15
* Cant A: 5
* Cant B: 3

S de positivo= -[5/7 \* log2(5/7) + 2/7 \* log2(2/7)] = 0.86

**Negativo=**

* Casos: 8
* Prop al nodo sup: 8/15
* Cant A: 5
* Cant B: 3

S de negativo= -[5/8\* log2(5/8) + 3/8 \* log2(3/8)] = 0.95

**S X3= 7/15\*0.86+8/15\*0.95= 0.90**

**Comparación:**

**S X1= 0.90**

**S X2= 0.85**

**S X1= 0.90**

**Conviene que el primer nodo empiece con la variable X2**