PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

APLICACIONES DE CIENCIAS DE LA COMPUTACION (INF265)

Examen 2

(Segundo Semestre 2023)

Indicaciones generales:

- Duración: 2 horas 45 minutos.
- No está permitido el uso de equipos electrónicos.
- **La presentación, la ortografía y la gramática influirán en la calificación.**
- **Sobre plagio:** Al encontrarse exámenes muy similares se procederá anular dichos exámenes, comunicando previamente a los involucrados.

Puntaje total: 20 puntos

Pregunta 1 (3 puntos)

Usted ha sido contratado por una entidad bancaria y le han encargado la construcción de un modelo de predicción de Riesgo de Impago. Para ello le han proporcionado un dataset de 100 registros de clientes con diferentes atributos y la variable target (Riesgo_Impago). Aquí la información que se muestra con la función .info()

```
clientes.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID                 100 non-null    int64
 1   Nombres            100 non-null    object
 2   Apellidos          98 non-null     object
 3   Direccion          98 non-null     object
 4   Sexo               97 non-null     object
 5   Edad               100 non-null    int64
 6   Estado_civil       100 non-null    object
 7   Educacion          100 non-null    object
 8   Tipo_empleo        100 non-null    object
 9   salario            99 non-null     float64
 10  Antigüedad_empleo  11 non-null     float64
 11  Riesgo_Impago      100 non-null    object
dtypes: float64(2), int64(2), object(8)
memory usage: 9.5+ KB
```

Estos son los detalles de cada variable:

ID: identificador único del cliente; **Nombres, Apellidos, Direccion**: nombres, apellidos y dirección del cliente (strings); **Sexo**: carácter que identifica el género del cliente (M:masculino, F:femenino); **Edad**: edad del cliente en años; **Estado_civil**: string del estado civil del cliente ('Soltero','Casado','Conviviente','Divorciado'); **Educacion**: nivel educativo del cliente ('Primaria','Secundaria','Instituto','Universidad'); **Tipo_empleo**: dependiente, independiente, negocio_propio; **salario**: salario del cliente; **Antigüedad_empleo**: antigüedad de empleo (en años); **Riesgo_Impago**: riesgo de impago del cliente ('bajo', 'moderado', 'elevado')

Plantee una secuencia de pasos de pre-procesamiento para obtener un conjunto de datos listo para ser usado para el entrenamiento de un modelo de machine learning. Para responder, use como plantilla la tabla de abajo, indicando el paso de pre-procesamiento, su comando en Python y la justificativa de por qué es necesario ese paso. Agregue tantos pasos como crea conveniente. El orden de pasos es importante.

Orden	Paso de procesamiento	Comando en Python	Justificativa
1			
2			

Pregunta 2 (2 puntos)

Explique brevemente qué es el problema de la "maldición de la dimensionalidad" en el contexto de *machine learning* y qué enfoques conoce para abordarlo

Pregunta 3 (2 puntos)

Explique qué es y para qué sirve el Análisis de Componentes Principales

Pregunta 4 (2 puntos)

Explique qué es y para qué sirve la regularización en regresión lineal. Qué tipos de regularización conoce y en qué se diferencian

Pregunta 5 (2 puntos)

En el algoritmo KNN, explique cómo influye el parámetro k en el subajuste/sobreajuste del modelo y por qué es importante escalar o estandarizar los datos para este algoritmo

Pregunta 6 (2 puntos)

En el algoritmo Regresion Tree, explique qué relación tiene el parámetro max_depth con el overfitting

Pregunta 7 (2 puntos)

Para qué sirve la medida de impureza en la construcción de un árbol de clasificación y qué medidas conoce. gini vs entropy

Pregunta 8 (2 puntos)

Cómo afecta la taza de aprendizaje en el algoritmo *back-propagation,* en cuanto a la velocidad de aprendizaje y convergencia de una red neuronal

a a a a
a b b a
orden de procesamiento

Pregunta 9 (2 puntos)

En una red neuronal Multilayer Perceptron (MLP) para clasificación es usual colocar como función de activación de las neuronas de salida la función SOFTMAX. Indique qué hace esta función y por qué es necesario colocarla en la capa de salida.

Pregunta 10 (1 punto)

Indique si es verdadero (V) o falso (F) y justifique su respuesta:

1. El algoritmo DBSCAN tiene 3 parámetros importantes: ***min_samples***, ***eps*** y el número de clusters ***k*** a encontrar	
2. El algoritmo K-means tiene la capacidad de detectar automáticamente la cantidad de clusters, a diferencia de DBSCAN	

Profesor del curso responsable del examen: **Edwin Villanueva Talavera**

Lima, 01 de diciembre de 2023