Seminario de tecnología **Nombre:** _____
2 Cuatrimestre, 2014
Examen Parcial 3
14/11/14
Tiempo: 120 minutos

Este examen consta de 5 paginas y 15 preguntas. Verifique que tiene todas las hojas necesarias. Las preguntas se responden en la misma hoja del examen.

El examen se puede realizar a libro abierto y esta permitido el uso de calculadora si es requerido. Se puede utilizar computadora.

Las siguientes reglas aplican para la aprobación del examen:

- Escritura de todas las **respuestas en tinta, sin excepción**.-
- **Se requiere un mínimo de 10 puntos** para la aprobación del examen.-
- Justificar sus respuestas, en caso de ser necesario, con diagramas o ejemplos claro.-
- Lea todo el examen antes de comenzar a responder. Algunas preguntas guardan relación con otras y pueden servir de ayuda.-

No escriba en la tabla de la derecha.

Mucha suerte! :)

Pregunta	Points	Score
1	2	
2	2	
3	1	
4	2	
5	1	
6	1	
7	1	
8	1	
9	1	
10	2	
11	1	
12	2	
13	1	
14	1	
15	1	
Total:	20	

1. (2 points) Defina el concepto de Minería de datos.

2. (2 points) Defina el concepto de cluster de datos. A que se refiere el concepto de clusterización?

3. (1 point) Indique cual/es de las siguientes herramientas se utilizan para análisis estadístico de clusters:

   - SPSS
   - Weka
   - Qucs

4. (2 points) **Indicar verdadero o falso**. Se desea crear un modelo regresión lineal a partir de un data set de entrenamiento. Agregar variables al modelo original siempre reducirá la suma de los cuadrados residuales medidos en el set de validación? **Justificar**.

5. (1 point) **Indicar verdadero o falso**. Aunque la selección hacia adelante (forward) y la eliminación hacia atrás (backward) son métodos rápidos para la selección de subconjuntos en regresión lineal, solo la selección por etapas garantiza encontrar el mejor subconjunto. **Justificar**.

6. (1 point) **Indicar verdadero o falso**. Un grupo de funciones de clasificación se ordenan utilizando el análisis discriminante para un conjunto de datos con tres clases $C1$, $C2$ y $C3$. Se asume que las tres clases son igualmente propensas a surgir en la aplicación. Posteriormente, se conoce que la probabilidad de $C1$ es el doble que la de $C2$ y $C3$. Las probabilidades para $C2$ y $C3$ son iguales. Si se vuelven a calcular las funciones de clasificación utilizando esta información, el valor de la función de clasificación para $C1$ se incrementará para cada punto de datos. **Justificar**.

7. (1 point) **Indicar verdadero o falso**. La tasa de errores de clasificación de un modelo de clasificación en el conjunto de validación es la mejor medida de la capacidad predictiva del modelo en los nuevos datos, a diferencia de su tasa de errores de clasificación en el conjunto de entrenamiento. **Justificar**.

8. (1 point) **Indicar verdadero o falso**. Un clasificador de redes neuronales para dos clases construye un límite de separación entre las clases que es lineal en sumas ponderadas de los valores de entrada. **Justificar**.

9. (1 point) Un conjunto de datos de 1.000 casos fue dividido en un conjunto de entrenamiento de 600 casos y un conjunto de validación de 400 casos. Un **k-Nearest Neighbors** con k = 1 tiene una tasa de error de clasificación del 8% sobre los datos de validación. Posteriormente se descubrió que la división se había hecho de forma incorrecta y que 100 casos del conjunto de datos de entrenamiento se había duplicado y accidentalmente había sobreescrito 100 casos en el conjunto de datos de validación. ¿Cuál es la tasa de error de clasificación para los 300 casos que estaban verdaderamente parte de la validación de datos?

10. (2 points) Describa las etapas del proceso de generación de un modelo de minería de datos.

11. (1 point) **Indique la opcion correcta.** El modelo de minería de datos que crea un algoritmo a partir de los datos puede tomar diversas formas, incluyendo:

    1. Un conjunto de clústeres que describe cómo se relacionan los casos de un conjunto de datos.
    2. Un árbol de decisión que predice un resultado y que describe cómo afectan a este los distintos criterios.
    3. Un modelo matemático que predice las ventas.
    4. Un conjunto de reglas que describen cómo se agrupan los productos en una transacción, y las probabilidades de que dichos productos se adquieran juntos.
    5. Ninguno de los anteriores
    6. Los itemas 1, 2, 3 y 4.

12. (2 points) Los algoritmos para el analisis de datos se pueden elegir, entre otros cosas, segun el tipo de analisis que se quiere realizar. Describa esta clasificacion.

13. (1 point) **Indique la opcion correcta.** La configuración de una estructura de minería de datos consta de 5 pasos. Marque cual de ellos es opcional.

    - Definir un origen de datos.
    - Seleccionar las columnas de datos que se van a incluir en la estructura (no es necesario agregar todas las columnas al modelo) y definir una clave.
    - Definir una clave para la estructura, incluyendo la clave de la tabla anidada, si procede.
    - Especificar si los datos de origen se deben separar en un conjunto de entrenamiento y en un conjunto de prueba.
    - Procesar la estructura.

14. (1 point) Defina el concepto de outlier. Porque es importante preprocesar la informacion y encontrar este tipo de datos?

15. (1 point) Existen 3 (tres) criterios para validar los modelos de minería de datos. Cuales son? Describalos.