Patrón de fondo

Descripción generada automáticamente con confianza media

Universidad Austral - Maestría en Ciencia de Datos

Trabajo Practico #1

**Autor: Andres Montes de Oca**

Fecha: 2 de Noviembre de 2022

Introduccion al

Data Mining

1. Que es el Data Mining?:

Es el proceso por el cual se extrae información valiosa desde fuentes de datos crudas o ‘raw data’. Este proceso utiliza distintas técnicas Estadística y de Machine Learning las cuales transforman la data cruda en ‘insights’ importantes para la toma de decisiones basadas en datos. Definitivamente no es una moda, yo diría que es mas un **Nuevo Paradigma**, muchos autores consideran al Data Mining o a la Ciencia de Datos como la **3ra Revolución Industrial**. Hoy por hoy casi todas las acciones que hacemos en nuestro día a día dejan una huella digital en algún lado. Esto nos abre un sinfín de oportunidades mediante el estudio y análisis de estos logs. Los pasos del proceso de Data Mining son:

1. Juntar la Data
2. Preprosesarla (encargarse de los nulls, los valores faltantes, outliers, etc)
3. Realización de un EDA (Exploratory Data Analysis) por el cual entenemos mejor de que se tratan los datos,
4. Generación de modelos estadísticos o machine learning, los cuales nos van a ayudar en la toma decisiones futuras
5. Prueba de performance y re rendimiento del modelo
6. Puesta en producción
7. Mantenimiento
8. Justifique V/F:

o Dividir los clientes de una compañía de acuerdo con su género. **Falso**

Esto parece una tarea de reporting, ya que los géneros deberían estar indicados en los datos

o Dividir los clientes de una compañía de acuerdo con su rentabilidad **Verdadero**

Si la rentabilidad no esta binarizada, esto podría ser una tarea de Segmentación (clustering) para el descubrimiento de nuevos grupos desconocidos hasta el momento

o Calcular el total de ventas de una compañía **Falso**

Esto también parece reporting

o Predecir los resultados de arrojar un par de dados. **Falso**

Esto me parece mas una tarea Estadística, y esta relacionada con la Esperanza del experimento

Predecir el precio futuro de las acciones de una compañía usando registros **Verdadero**

Esto parece ser un modelo de regresión lineal de Machine Learning Supervisado

o Monitorear los latidos del corazón de un paciente para buscar anomalías. **Falso**

No parece serlo, es una tarea de monitoreo medico

o Monitorear ondas sísmicas para detectar actividades de terremotos. **Verdadero**

Analizando datos históricos se puede generar algún modelo de clasificación el cual nos advierta sobre la potencial aparición de un sismo

o Extraer las frecuencias de una onda de sonido. **NS/NC**

1. Ejemplom Data Mining:

En las áreas de RRHH de las grandes empresas, en análisis de churn es un buen ejemplo. Con el se busca predecir que empleados pueden potencialmente dejar la compañía por busca de oportunidades en otras empresas. Para esto es fundamental contar con varias variables de información sobre mis empleados, la calidad del dato es fundamental

1. Ejemplo consultor:
2. Importancia sobre la privacidad de los datos:

* Datos censales del período 1980-1998:

No me parece que la privacidad sea importante aquí ya que son datos públicos

* Direccions IP y tiempos:

En este caso si es importante, ya que por la IP se pude llegar a geolocalizar a los usuarios conectados al sitio web

* Imagnes Satelitales:

Excluyendo las zonas con protección militar, me parece que no habría problemas de privacidad aquí

* Nombres y direcciones de email:

Definitivamente si, estos son datos privados

1. CRISP-DM vs SAS

CRISP-DM tiene una fase adicional al comienzo, relacionada con el entendimiento de los requerimientos de negocio, y otra fase adicional al final relacionada con la implementación o puesta en producción de mi modelo. Pero ambas metodologías también tienen varias etapas en común, como la recolección de datos con sus EDAs, transformaciones de variables, selección de el/los modelos a utilizar, y sus evaluaciones de performance