

Tarea 1: Preprocesamiento de datos

F.C/W.G GD M.D - II-2015

Becas Crema

Objetivo:

La Organización de Becas Crema lo ha contratado para que los ayude a obtener una vista minable basada en las encuestas de los estudiantes que acaban de renovar su afiliación a la ayuda económica.

Para discernir las características de una ‘buena’ vista minable puede usted referirse a: [Tidy Data](#)

Requerimientos:

Documento de explicación del set de datos:

La Organización perdió la explicación del las características del dataset **data.csv**.

Como es de esperar, el dueño no solo necesita la vista minable sino una explicación de cada característica en donde debe incluir información necesaria para entender cada una de las columnas del dataset.

Es tarea de usted inferir el contexto, de poderse, para explicar la pregunta aunado a agregar información adicional necesaria que ayude al dueño a entender su vista minable.

Este documento lo llamará **explicacion.pdf**. Es de suma importancia impresionar al dueño por lo que es importante la presentación del mismo.

Proceso de generación de la vista minable

El “**proceso**” no es más que ejecutar un script en python que genere como salida la vista minable. La entrada es el archivo **data.csv** y la salida **minable.csv**.

No existe un proceso único para generar la vista minable. En consecuencia, queda a juicio del experto (usted) el uso de técnicas de preprocesamiento y limpieza de datos para realizar su trabajo. Entre los aspectos a tomar en cuenta tenemos:

1. Eliminar instancias o características usando algún criterio visto en clase.
2. Transformar los tipos de datos iniciales de cada columna según su criterio.
3. Inputar valores en las columnas que lo requieran.
4. Señalar los valores anómalos.
5. Aplicar estandarización, normalización y las técnicas disponibles en scikit-learn en columnas que así lo requiera según su criterio.
6. Usar métricas de centralidad.
7. Utilizar alguna técnica para ajustar la dimensionalidad del set de datos equilibrando la menor redundancia y lamayor cantidad de información.
8. Proveer las condiciones iniciales/finales de los datos(vistas en clase y que considere necesarias).

Se anima y se da valor al uso de técnicas no mencionadas anteriormente. Se recompensará con puntos extras.

Documento de explicación de la vista minable

En este documento **informe.pdf** debe explicar todas las decisiones que tomó en cuenta a la hora de realizar la vista minable. Esto debe incluir un análisis exploratorio de los datos de las características y, de ser necesario, utilice gráficas para ayudar su análisis (se recomienda usar el paquete **matplotlib** de Python).

Consideraciones de forma:

Ingresa a la dirección [DataMiningUCV](#) y siga las instrucciones para crear un repositorio en [GitHub](#) perteneciente a la organización **DataMiningUCV**.

Este repositorio será propiedad de la organización pero solo usted puede realizar cambios en el mismo. El repositorio debe poseer lo siguiente:

1. Script (.py) intradocumentado.
2. Propuesta de vista minable en un archivo **minable.csv**.
3. README.md explicando la configuración del ambiente en el cual trabajó:
 - Ejemplo: [README.md de Bootstrap](#)
4. Explicación de la vista minable **informe.pdf**.
5. Explicación del documento de explicación del set de datos **explicacion.pdf**.

Consideraciones de contenido:

- Se recomienda el uso de las funcionalidades de preprocesamiento provista por el paquete [scikit-learn](#), así como el uso de las funcionalidades del paquete [pandas](#).
- La tarea es **estrictamente** individual. Se promueve la participación y discusión de la misma en un ambiente responsable. Sin embargo, cualquier evidencia de copia será severamente sancionada colocando una nota mínima de cero (0) puntos según lo establecido en la Ley de Universidades. **Cualquier proyecto entregado** debe ser fruto de su propio trabajo.
- Fecha de Entrega: **Viernes 15 de Enero de 2016**.
 - Hasta este día se aceptarán push's en los repositorios.
 - No se recibirá ninguna tarea por correo electrónico.