Universidad del País Vasco Euskal Herriko Unibertsitatea

Sistemas de Ayuda a la Decisión

Práctica 1

Índice

1. Objetivos y Descripción de los contenidos

Esta práctica, junto con la práctica 2, tiene como objetivo adquirir las competencias que serán evaluadas en la primera prueba práctica individual que tendrá lugar hacia la cuarta semana. Por lo tanto esta tarea se enmarca dentro de la evaluación continua pero no dentro de los hitos evaluables.

Las **competencias** que el alumno deberá haber adquirido tras realizar la práctica son:

1. Capacidad para crear un proyecto, importar datos en el y explorar y analizar los datos importados y su calidad.

2. Capacidad para describir y caracterizar los datos. Cuantitativos (numéricos) o Categóricos (ordinales o nominales), media, mediana y la moda, histogramas versus diagramas de barras, diagramas de cajas, valores faltantes (missing values), valores atípicos (outliers) etc.

3. Capacidad para realizar un preprocesado mínimo generando recetas de python para tratar valores faltantes, escalar y discretizar.

2. Materiales disponibles

El siguiente material será empleado para el desarrollo de las tareas que se proponen y que tiene como objetivo alcanzar las competencias enumeradas anteriormente.

1. Datos: Los aportados por DSS para los tutoriales 101 y 102.
2. Tutoriales: Se pone a disposición del alumno los siguientes tutoriales de libre acceso tras previo registro[1]:
   - https://academy.dataiku.com/basics-101
3. Pequeños vídeos con píldoras mostrando como generar recetas python (http://lsi.bp.ehu.es/asignaturas/SAD-2021-2022/practica1/)
4. Recetas de Python que se encuentran en eGela
5. Lecturas: Se encontrarán en eGela las lecturas asociadas a esta práctica y que son las que previamente se han trabajado en la sesión teórica (Capítulos 1.1, 1.3, 1.6 del libro de Weka).
6. Transparencias: Se encontrarán en eGela la transparencias asociadas que recogen el preprocesado de los datos.
7. Tutorial de pandas: https://www.kaggle.com/learn/pandas (tiempo estimado 4 horas)

3. Tareas

Para obtener los objetivos buscados en este proyecto se proponen las siguientes **tareas**:

- Visualizar lod tutoriales.
- Responder a preguntas sobre el mismo.
- Repetir los pasos presentados en el tutorial de una manera guiada, de forma autónoma sobre un nuevo conjunto de datos.
- Repetir los pasos presentados en las píldoras y generalizarlos para todos los campos (_FORMAT)

[1] Para aquellos que quieran ver los subtítulos en castellano podéis acceder al tutorial https://www.youtube.com/watch?v=T2pmh578Nac&list=PLWjlCkA2BrRQ6Aro34zIVzZ_uhdK7eJ4m&index=1 y en la configuración (la ruedita) seleccionar que se traduzcan los subtítulos a castellano

3.1. Tarea previa guiada 1: Visualización del tutorial de exploración de los datos.

Visualizar el tutorial : https://academy.dataiku.com/basics-101
Duración total de los vídeos cortos aprox. **30 min**
Este tutorial trabaja los siguientes conceptos:

- Crear un proyecto

  Concepto: Proyecto

  Tutorial práctico: Crear el proyecto

  Prueba: Crear un proyecto

- Crear un Conjunto de Datos

  Importar Conjunto de datos

  Concepto: Conjunto de datos

  Concepto: Conexiones

  Tutorial práctico: Crear el conjunto de datos

  Prueba: Crear el conjunto de datos

- Explorar los datos

  Concepto: Esquema

  Concepto: Tipo de almacenamiento

  Concepto: Significado

  Concepto: Muestreo

  Conceptos: Análisis y Calidad de los datos

  Concepto: Gráficos

  Tutorial práctico: exploración de datos

Cuestionario:
Duración aprox. **10 min.**

1. Pregunta: Al diseñar el flujo, como cuando se usa un script de receta Preparar para cambiar el tipo de almacenamiento o el significado de una columna, Dataiku DSS le pregunta si desea actualizar el esquema.
   - Verdadero. Esto se debe a que el esquema del conjunto de datos de salida cambia a medida que aplica cambios a las columnas.
   - Falso. Solo los cambios en el tipo de almacenamiento de una columna harán que DSS le pregunte si desea actualizar el esquema.
2. Pregunta: Selecciona la mejor respuesta para completar el espacio en blanco: String, integer, float y boolean son ejemplos de _____ en Dataiku DSS.

- Extensiones de archivo
- Estos pueden ser tipos de almacenamiento o significados.
- Tipos de almacenamiento
- Significados (meaning)

3. Pregunta: ¿Cuál de los siguientes tipos usaría para asignar a una columna un tipo semántico rico, como *GeoPoint* o *Gender*?

   - Tipo de almacenamiento
   - Tanto el tipo de almacenamiento como el significado (meaning) pueden lograr esto.
   - Ninguno de éstos.
   - Significado (meaning)

4. Pregunta:

   - Todos estos son ejemplos de tipos de almacenamiento de columnas en Dataiku, ¿EXCEPTO?
   - booleano
   - Float
   - Entero
   - Nombre (Name)

5. Pregunta: Para encontrar valores atípicos (outliers) en los valores de la columna de un conjunto de datos, debe usar la pestaña Gráficos.

   - Falso: puede utilizar la ventana Analizar.
   - Verdadero: debe crear un gráfico.

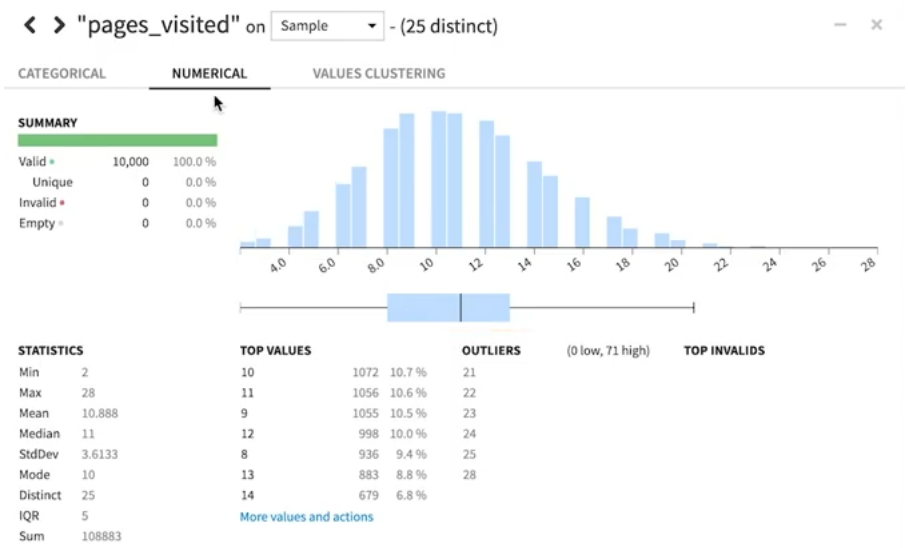
6. Pregunta: ¿Cuáles de las siguientes afirmaciones son verdaderas sobre el muestreo en DSS? (Escoge dos).

   - El muestreo permite a los usuarios trabajar de forma interactiva con grandes conjuntos de datos.
   - Los ajustes de muestreo son configurables.
   - Al preparar un conjunto de datos o crear un gráfico, Dataiku usa todas las filas del conjunto de datos de forma predeterminada.

7. Pregunta: Empareja los términos de la izquierda con sus descripciones a derecha.

8. Pregunta: Empareja cada término de la ventana de análisis con su definición (los términos y sus correspondencias se encuentran al final del documento).

1. Primeros N items	Resp-A. Selecciona del conjunto completo de datos N de forma aleatoria para generar la muestra
2. Primeros N items aleatorios	Resp-B. Selecciona de forma secuencial los N primeros items del conjunto total de datos para generar la muestra
3. Primeros N items aleatorios estratificados	Resp-C. Selecciona del conjunto completo de datos N de forma aleatoria tratando de mantener el mismo número de items o instancias por cada clase y así generar la muestra
4. Primeros N items aleatorios estratificados y balanceados	Resp-D. Selecciona del conjunto completo de datos N de forma aleatoria manteniendo la proporción de items o instancias del conjunto completo por cada clase y así generar la muestra. Es decir, la muestra mantiene la distribución de clases del conjunto origen (el que contiene todas las instancias).



3.2. Tarea autónoma 1: Creación de un nuevo proyecto, importación de los datos, análisis, exploración y caracterización de los datos.

¡Ahora es la hora de la verdad! El alumno deberá demostrarse a sí mismo que ha adquirido las competencias esperadas. Para ello se le solicita que cree un nuevo proyecto, que importe los datos almacenados en el fichero ACC_AUX2_labo.csv, que explore los datos contenidos en el, y que los caracterice. Así, el alumno deberá de ser capaz de dado un atributo obtener la ventana de análisis, que le mostrará la información necesaria para responder a las preguntas que se le plantean al final y que le permitirán medir sus logros en lo referente a las competencias descritas al comienzo de este documento.

Para llevar a cabo esta **subtarea** el alumno deberá bajar del servidor el fichero ACC_AUX2_labo.csv que encontrará en eGela. Este fichero pertenece al conjunto de datos CarCrashFARS2019-Dataset y contiene información acerca de los accidentes sucedidos entre los años 1982 a 2019 en Estados Unidos.
Duración: Aprox. **20 min**.

La siguiente tabla describe algunos de los atributos que aparecen recogidos en ese fichero.

Accident File (ACC_AUX)

Variable	Description
A_CRAINJ	Crash Injury Type
A_CT	Crash Type
A_D15_19	Crashes Involving a Young Driver (Aged 15-19)
A_D15_20	Crashes Involving a Young Driver (Aged 15-20)
A_D16_19	Crashes Involving a Young Driver (Aged 16-19)
A_D16_20	Crashes Involving a Young Driver (Aged 16-20)
A_D16_24	Crashes Involving a Young Driver (Aged 16-24)
A_D21_24	Crashes Involving a Young Driver (Aged 21-24)
A_D65PLS	Crashes Involving an Older Driver (Aged 65+)
A_DIST	Involving a Distracted Driver
A_DOW	Day of Week
A_DROWSY	Involving a Drowsy Driver
A_HR	Involving a Hit and Run
A_INTER	Interstate
A_INTSEC	Intersection
A_JUNC	Junction
A_LT	Involving a Large Truck
A_MANCOL	Manner of Collision
A_MC	Involving a Motorcycle
A_PED	Involving a Pedestrian
A_PEDAL	Involving a Pedalcyclist
A_POLPUR	Involving a Police Pursuit
A_POSBAC	Involving a Driver with a Positive BAC Test Result
A_RD	Involving a Roadway Departure (FHWA definition)
A_REGION	NHTSA Region
A_RELRD	Relationship to the Trafficway
A_ROADFC	Roadway Function Class
A_ROLL	Involving a Rollover
A_RU	Land Use (Rural/Urban)
A_SPCRA	Involving Speeding
A_TOD	Time of Day
BIA	Tribal lands based on geographic location and spatial data
SPJ_INDIAN	Special Jurisdiction Indian Reservation
INDIAN_RES	Indian Reservation based on special jurisdiction and geographic location data

Responde a las siguientes preguntas:

1. Tomando los atributos STATE y STATE_FORMAT ¿Cúal es un atributo categorial y cúal es numérico discreto?
2. ¿Qué atributos crees que son redundantes?
3. Se pide realizar 3 tipos de muestreos y posteriormente responder a unas preguntas.
   a) Muestreo de las primeras 10.000 instancias en el orden secuencial
   b) Muestreo de 10.000 instancias de forma random.
   c) Muestreo de 10.000 instancias de forma estratificada en base al atributo STATE.

Responder a las siguientes preguntas (cuando sea posible) para cada uno de los muestreos. Para ello nos centraremos en el análisis de dos atributos, FATALS, STATE y STATE_FORMAT:

- ¿Cúal es el número de instancias?
- ¿Cuantos valores distintos hay?
- ¿Cuantos valores únicos hay?
- ¿Cúales son los valores max. y min.?
- ¿Hay algún valor atípico?

Repetid la tarea modificando el tamaño de la muestra de 10.000 a 1.000 instancias.

3.3. Tarea previa guiada 2: Visualización del tutorial de transformación de los datos.

Visualizar el tutorial : https://academy.dataiku.com/basics-102
Duración total de los vídeos cortos aprox. **30 min**
Este tutorial trabaja los siguientes conceptos:

- Preparar los datos creando recetas
  - Concepto: Receta
  - Concepto: preparar receta
  - Concepto: Manejo de fechas
  - Concepto: Fórmula
  - Tutorial práctico: Preparar los datos
- Crear estadísticas visuales interactivas
  - Concepto: Hoja de trabajo de estadísticas
  - Concepto: Tarjeta de estadísticas
  - Tutorial práctico: Obtener estadísticas visuales interactivas
  - Prueba: Obtener estadísticas visuales interactivas
- Agrupar los datos
  - Concepto: Receta grupal
  - Tutorial práctico: agrupe los datos
  - Prueba: Agrupar los datos
- Exploración del flujo

Concepto: Flujo

Concepto: Computation Engine

Concepto: Vista de trabajo

Tutorial práctico: Explore el flujo

Prueba: Explore el flujo

Cuestionario:
Duración aprox. **10 min.**

1. Pregunta: ¿Cuál de los siguientes tipos de expresiones podría usar en un paso de fórmula de Dataiku? Seleccione todas las que correspondan
   - Operadores lógicos: AND, OR
   - Pruebas para valores perdidos: isBlank(), isNULL()
   - Operadores de comparación:>, <, >=, <=
   - Funciones matemáticas comunes: redondear, suma, máx.
2. Pregunta: ¿Cuáles son algunas formas en las que puede agregar pasos al guión en una receta de preparación? (Selecciona todas las que correspondan.)
   - Seleccionar los pasos sugeridos en el menú contextual de una columna.
   - Agregar pasos a través de la ventana Analizar
   - Agregar un paso directamente desde la biblioteca del procesador
   - Arrastra las columnas para ajustar su orden.
3. Pregunta: Seleccionar aquellas que se considere verdaderas.
   - Cuando la infraestructura de almacenamiento subyacente del conjunto de datos cambia, afecta la lógica de procesamiento que se encuentra en las recetas de un flujo.
   - Los programadores pueden definir su propia lógica de procesamiento en una receta de código.
4. Pregunta: Una vez que se ha agregado un paso a una secuencia de comandos Preparar receta, DSS aplica inmediatamente ese paso al conjunto de datos de entrada completo.
   - Verdadero. El paso se aplica inmediatamente al conjunto de datos de entrada completo, pero puede deshabilitar la vista previa del paso si tarda demasiado.
   - Falso. El paso se aplica inmediatamente a la muestra del conjunto de datos actual, lo que permite una retroalimentación visual inmediata. El paso no se aplica al conjunto de datos de entrada completo hasta que se ejecuta la receta.

5. Pregunta:¿Qué opción posibilita agrupar el conjunto de datos de acuerdo a una variable específica de modo que pueda realizar cálculos en cada subgrupo de datos?
   - Agrupar por menú
   - Dividir por menú
   - Menú de configuración
   - Menú de muestreo
6. Pregunta: Si un conjunto de datos de seis filas con tres clientes distintos está agrupado por clientes, el conjunto de datos de salida también contendrá seis filas.
   - Verdadero. La receta de grupo mantiene las dimensiones originales del conjunto de datos de entrada y agrega las agregaciones al conjunto de datos.
   - Falso. La receta de grupo reduce el conjunto de datos de entrada al número de valores distintos en la clave de grupo. Por ejemplo, en el escenario anterior, el conjunto de datos de salida solo contendría tres filas.

3.4. Tarea guiada autónoma 3: Visualización de las píldoras de escalado, transformación de atributos categoriales a numéricos y tratamiento de los valores faltantes.

Tarea:
Duración aprox. **90 min.**

- Visualizar las pi¡íldoras (aprox. 20 min). http://lsi.bp.ehu.es/asignaturas/SAD-2021-2022/practica1/
- Repetir los pasos que aparecen en las pildoras.
- Probar los 3 métodos de escalado (z-score, min-max, max).
- No hay pildora para la última receta que aparece en el fichero de recetas. Probar la última receta en Dataiku.

Referencias

[1] https://crashstats.nhtsa.dot.gov/#!/DocumentTypeList/23

[2] https://academy.dataiku.com/basics-101

[3] https://academy.dataiku.com/basics-102

[4] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco, 2000

[5] https://www.kaggle.com/learn/pandas

[6] Competencias asociadas y estructura de la tarea: Se han tomado como referencia los apuntes de SAD 2020-2021 (fuente: Alicia Pérez) para así coordinar que la práctica y las competencias a obtener sea lo más similar posible a lo solicitado en años anteriores.

1. Media	Resp-A. Número de veces en los que el atributo objeto de análisis aparece vacio. Recordad que los datos no son perfectos en la mayoría de los casos. En muchas ocasiones algunos items o instancias carecen de valor para algún atributo.
2. Mediana	Resp-B. Representa el valor más frecuente. Si todos los valores fuesen únicos no existiría la moda.
3. Moda	Resp-C. En un histograma que es el diagrama que visualiza la distribución delos datos, son valores extremos que aparecen en los extremos de los histogramas. Por definición son datos poco frecuentes de ahí su anomalía con respecto al resto de los valores.
4. Outliers (valores atípicos)	Resp-D. Busca representar donde se encuentran centrados los datos y se emplea para representar el valor central cuando los datos contienen muchos valores atípicos o extremos. Es uno de los valores que aparece representado en el box plot o diagrama de caja junto con los cuartiles.
5. Empty or missing (valores faltantes)	Resp-E. Número de instancias que contiene la muestra empleada por DSS para hacer el análisis de los datos.
6. Valores únicos	Resp-F. Número de valores distintos que toma un atributo.
7. Valores distintos	Resp-G. Busca representar donde se encuentran centrados los datos y se emplea para representar el valor central cuando los datos no contienen muchos valores atípicos o extremos. Se denomia también promedio.
8. Tamaño de la muestra	Resp-H. Número de valores únicos, es decir, valores para el atributo objeto de análisis que se muestran con frecuencia 1.

Cuadro 1: Pares de término definición asociados a la pregunta 8