Práctica 2 (35% nota final) Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Para hacer esta práctica tendréis que trabajar en grupos de 3 o 2 personas, o si preferís, también podéis hacerlo de manera individual. Tendréis que entregar un solo archivo con el enlace Github (https://github.com) donde haya las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github.

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Las diferentes tareas a realizar (y **justificar**) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?
2. Limpieza de los datos.
   2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?
   2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?
3. Análisis de los datos.
   3.1. Selección de los grupos de datos que se quieren analizar/comparar.
   3.2. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.
   3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.
4. Representación de los resultados a partir de tablas y gráficas.
5. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
6. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

- Megan Squire (2015). *Clean Data*. Packt Publishing Ltd.
- Jason W. Osborne (2012). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. SAGE Publications, Inc.
- Peter Dalgaard (2008). *Introductory statistics with R*. Springer Science & Business Media.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- El apartado 1 vale 0,5 puntos.

- Los apartados 2 y 5 valen 1 punto.
- Los apartados 3, 4 y 6 valen 2,5 puntos.

Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Las diferentes etapas deberán justificarse y acompañarse del código correspondiente. También se valorará la síntesis y claridad, a través del uso de comentarios, del código resultante, así como la calidad de los datos finales analizados.

Formato y fecha de entrega

Durante la semana del 11 de diciembre el grupo podrá entregar al profesor una entrega parcial opcional. Esta entrega parcial es muy recomendable para recibir asesoramiento sobre la práctica y verificar que la dirección tomada es la correcta. Se entregarán comentarios a los estudiantes que hayan efectuado la entrega parcial pero no contará para la nota de la práctica. En la entrega parcial los estudiantes deberán entregar por correo electrónico (mcalvogonza@uoc.edu) el enlace al repositorio Github con el que hayan avanzado.

En referente a la entrega final, hay que entregar un único fichero que contenga el enlace Github donde haya:

1. Una Wiki con los nombres de los componentes del grupo y una descripción de los ficheros.
2. Un documento Word, Open Office o PDF con las respuestas a las preguntas y los nombres de los componentes del grupo.
3. Una carpeta con el código generado para analizar los datos.
4. El fichero CSV con los datos originales.
5. El fichero CSV con los datos finales analizados.

Este documento de entrega final de la Práctica 2 se tiene que entregar en el espacio de Entrega y Registro de AC del aula antes de las **23:59** del día **8 de enero**. No se aceptarán entregas fuera de plazo.

Respuestas

1. El dataset seleccionado es https://www.kaggle.com/devinanzelmo/dota-2-matches . Este dataset corresponde a unos datos recogidos de un video juego que se llama Dota 2 (http://es.dota2.com/) este juego es un MOBA (https://es.wikipedia.org/wiki/Multiplayer_online_battle_arena). Estos juego combina grandes reflejos con un gran conocimiento de las estrategias del juego debido a su profundidad. Lo he elegido ya que tenia que hacer la practica poder hacerlo sobre algo que me gusta y me llama la atención. Se pretende resolver si los héroes que son de distancia (disparan o lanzan sus proyectiles) tienen más porcentaje de victoria que los que tienen que golpear cuerpo a cuerpo.
2. Los campos mas relevantes son seleccionar los héroes que son a distancia (https://dota2.gamepedia.com/Category:Ranged_heroes) y el win ratio en las partidas.
   En nuestro caso no hemos encontrado valores extremos, ya que los datos a analizar son sencillos y el dataset de partida esta muy depurado.
3. El grupo de dato a analizar son los players_sym y los match_sym.
4. NA
5. No he podido resolver el problema por no tener los conocimientos necesarios sobre R para resolver el problema que he planteado.
6. Como no he sido capaz de hacer un código limpio y funcional pongo el “History” que me proporciona R-Studio; como muestra del trabajo que he intentado llevar acabo:

```
hero_names <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/hero_names.csv", sep=";")
View(hero_names)
hero_names <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/hero_names.csv", sep=";")
View(hero_names)
save.image("C:/Users/Akrasia/Desktop/PR2/.RData")
range_heroes <-(hero_names, Range == True)
range_heroes <-(hero_names, hero_names$Range == True)
range_heroes <- (hero_names, hero_names$Range == True)
range_heroes <- (hero_names , hero_names$Range == True)
hero_names <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/hero_names.csv", sep=";")
View(hero_names)
hero_names <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/hero_names.csv", sep=";")
View(hero_names)
```

```
range_heroes <- (hero_names, hero_names$Range == 1)
range_heroes <- (hero_names , hero_names$Range == 1)
range_heroes <- (hero_names , Range == 1)
range_heroes <- (hero_names, Range == 1)
range_heroes <-(hero_names, Range == 1)
range_heroes <- subset(hero_names, Range == 1)
View(range_heroes)
save.image("C:/Users/Akrasia/Desktop/PR2/.RData")
[79,3] = 1
[79,3] <- 1
[79,"Range"] <- 1
hero_names[79,3] = 1
range_heroes <-subset(hero_names, Range == 1)
save.image("C:/Users/Akrasia/Desktop/PR2/.RData")
melee_heroes <-subset(hero_names, Range == 0)
View(hero_names)
hero_names <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/hero_names.csv", sep=";")
View(hero_names)
save.image("C:/Users/Akrasia/Desktop/PR2/.RData")
View(range_heroes)
View(melee_heroes)
save.image("C:/Users/Akrasia/Desktop/PR2/.RData")
match <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/match.csv")
View(match)
match_outcomes <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/match_outcomes.csv")
View(match_outcomes)
View(match_outcomes)
players <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/players.csv")
View(players)
match <- read.csv("C:/Users/Akrasia/Desktop/dota-2-matches/match.csv")
View(match)
View(match)
View(players)
match_sym <- match[,c("match_id", "radiant_win")]
View(match_sym)
save.image("C:/Users/Akrasia/Desktop/PR2/.RData")
players_sym <- players[,1:4]
View(players_sym)
save.image("C:/Users/Akrasia/Desktop/PR2/.RData")
save.image("C:/Users/Akrasia/Desktop/PR2/.RData")
```

```
match_sym_exp <- macth_sym[rep(row(match_sym),9)]
players_sym["Range_heroes"]
players_sym["Range_heroes"] <- NA
View(range_heroes)
View(players_sym)
```