**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 2 personas. Tendréis que entregar un solo archivo con el enlace Github (https://github.com) donde se encuentren 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. Aunque no se trata del mismo enunciado, los siguientes ejemplos de ediciones anteriores os pueden servir como guía:

● Ejemplo: https://github.com/Bengis/nba-gap-cleaning

● Ejemplo complejo (archivo adjunto).

**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**

Este proyecto ha sido desarrollado y elaborado por:

- Mónica Alexandra Gómez Martínez

- David Francisco Dávila Ortega

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Algunos ejemplos de dataset con los que podéis trabajar son:

● Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)

● Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y **justificar**) son las siguientes:

**1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende**

**responder?**

El dataset elegido se llama ‘Credit Card customers’ y puede descargarse desde la plataforma

[Kaggle](https://www.kaggle.com/sakshigoyal7/credit-card-customers). Su importancia radica en que trata de resolver un frecuente problema de negocios.

Este conjunto de datos contiene la información de una agencia bancaria, donde el

administrador desea conocer el o los motivos por los cuales están perdiendo clientes, así

como prevenir el así llamado ‘Credit Card Churning’.

Habiendo identificado los clientes que potencialmente desean prescindir de algún servicio

bancario, la agencia podría anticipar estrategias para lograr que dichos clientes se queden

por más tiempo.

Este dataset contiene información sobre 10,127 clientes. Entre los datos descriptores se

encuentran 23 variables que son:

0 CLIENTNUM: de tipo entero y sin valores nulos

1 Attrition\_Flag: de tipo entero y sin valores nulos

2 Customer\_Age: de tipo entero y sin valores nulos

3 Gender: de tipo objeto y sin valores nulos

4 Dependent\_count: de tipo entero y sin valores nulos

5 Education\_Level: de tipo objeto y sin valores nulos

6 Marital\_Status: de tipo objeto y sin valores nulos

7 Income\_Category: de tipo objeto y sin valores nulos

8 Card\_Category: de tipo objeto y sin valores nulos

9 Months\_on\_book: de tipo entero y sin valores nulos

10 Total\_Relationship\_Count: de tipo entero y sin valores nulos

11 Months\_Inactive\_12\_mon: de tipo entero y sin valores nulos

12 Contacts\_Count\_12\_mon: de tipo entero y sin valores nulos

13 Credit\_Limit: de tipo flotante y sin valores nulos

14 Total\_Revolving\_Bal: de tipo entero y sin valores nulos

15 Avg\_Open\_To\_Buy: de tipo flotante y sin valores nulos

16 Total\_Amt\_Chng\_Q4\_Q1: de tipo flotante y sin valores nulos

17 Total\_Trans\_Amt: de tipo entero y sin valores nulos

18 Total\_Trans\_Ct: de tipo entero y sin valores nulos

19 Total\_Ct\_Chng\_Q4\_Q1: de tipo flotante y sin valores nulos

20 Avg\_Utilization\_Ratio: de tipo flotante y sin valores nulos

21 Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1: de tipo objeto y sin valores nulos

22 Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2: de tipo flotante y sin valores nulos

Este dataset se encuentra completamente en el idioma inglés y se trabajará como tal sin hacer traducciones de variables o información. Las modificaciones realizadas (preprocesamiento) se encuentran comentadas dentro del archivo Rmarkdown (PRA2\_Limpieza.Rmd).

**2. Integración y selección de los datos de interés a analizar.**

Para ello se trabajó con Rmarkdown en Rstudio:

**Selección de los datos de interés.-** Para el presente proyecto se utilizarán todas las variables presentes en el juego de datos a excepción de las dos últimas que no serán de utilidad en los análisis que se planean hacer a posteriori.

**Análisis gráfico descriptivo.-** De manera general se observa que las variables cuantitativas presentan una alta dispersión en sus datos, en cuanto a las variables cualitativas, por ejemplo, se puede observar que el banco en cuestión presenta una mayor cantidad de clientes con cuentas activas que inactivas, existen más mujeres que hombres, la mayoría están solteros o casados con estudios terminados y que usan la tarjeta de crédito categoría “Blue”.

**3. Limpieza de los datos.**

El proceso paso a paso se encuentra en el archivo: ‘PRA2\_Limpieza.html’, los resultados se encuentran en el archivo ‘BankChurners.csv’ y el dataset procesado se denomina ‘‘BankChurners\_clean.csv’

**3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?**

Este dataset en particular presenta una completitud del 100%. En el caso de haber

encontrado elementos vacíos o nulos es conveniente conocer la fuente de los datos

antes de proceder a eliminarlos ya que se puede estar perdiendo valiosa información.

Lo que se hubiera hecho en este caso es una imputación de valores vacíos/nulos y, en

medida de lo posible, evitar la eliminación de algún valor.

**3.2. Identificación y tratamiento de valores extremos.**

La distribución no es simétrica para la mayoría de variables, y es simétrica en los casos de: "Customer\_Age", "Dependent\_count" y "Months\_on\_book".

Para la mayoría de variables se ha optado por mantener los valores originales con excepción de aquellos outliers encontrados para las variables: “”Customer\_Age” y Total\_Trans\_Ct”.

Un análisis variable por variable se detalla en el archivo Rmarkdown (.Rmd).

**4. Análisis de los datos.**

**4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).**

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que

pueden resultar interesantes para analizar y/o comparar.

**4.2. Comprobación de la normalidad y homogeneidad de la varianza.**

**4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.**

**5. Representación de los resultados a partir de tablas y gráficas.**

**6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?**

**7. 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:

● Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

● Megan Squire (2015). *Clean Data*. Packt Publishing Ltd.

● Jiawei Han, Micheine Kamber, Jian Pei (2012). *Data mining: concepts and techniques*. Morgan Kaufmann.

● Jason W. Osborne (2010). *Data Cleaning Basics: Best Practices in Dealing with Extreme Scores.* Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

● Peter Dalgaard (2008). *Introductory statistics with R*. Springer Science & Business Media. ● Wes McKinney (2012). *Python for Data Analysis*. O’Reilley Media, Inc.

● 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:

* Los apartados 1, 2 y 6 valen 0,5 puntos.
* Los apartados 3, 5 y 7 valen 2 puntos.
* El apartado 4 vale 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 21 al 25 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, al profesor encargado del aula, 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, el cual no se podrá modificar posteriormente a la fecha de entrega, donde haya:

1. Una Wiki con los nombres de los componentes del grupo y una descripción de los ficheros.
2. Un documento PDF con las respuestas a las preguntas y los nombres de los componentes del grupo. Además, al final del documento, deberá aparecer la siguiente tabla de contribuciones al trabajo, la cual debe firmar cada integrante del grupo con sus iniciales. Las iniciales representan la confirmación de que el integrante ha participado en dicho apartado. Todos los integrantes deben participar en cada apartado, por lo que, idealmente, los apartados deberían estar firmados por todos los integrantes.

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa | **DFDO (dfdavila)**  **Mónica Alexandra Gómez Martínez** |
| Redacción de las respuestas | **DFDO (dfdavila)**  **Mónica Alexandra Gómez Martínez** |
| Desarrollo código | **DFDO (dfdavila)**  **Mónica Alexandra Gómez Martínez** |

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 debe entregar en el espacio de Entrega y Registro de AC del aula antes de las **23:59** del día **5 de enero**. No se aceptarán entregas fuera de plazo.