Materiales D.A promo C

Proyecto 2

Se va a tomar el dataset que se ha a utilizado para el proyecto del módulo 1, una vez aplicada la limpieza preliminar. Con estos archivos procesados vamos a montar de nuevo el dataset.

Índice

Resumen

En este proyecto vamos a aprender a tratar con los archivos que se han obtenido de el procesamiento de los ficheros del proyecto del módulo 1. Para ello tenemos que aprender a unificar diferentes fuentes de datos en un único fichero, aplicar la limpieza que nos parezca conveniente, transformar los datos y ser capaz de extraer conocimiento de la información contenida. Para ello vamos a tener que presentar un informe al final del módulo con los resultados obtenidos de forma visual y ser capaces de explicar algunas cosas interesantes del análisis realizado ¡Esta será vuestra segunda experiencia de trabajo en equipo relacionada con programación! ¿Estáis preparadas?

Objetivos

1. Consolidar los conocimientos de Numpy, Pandas, Matplotlib y Seaborn. Para procesar y unificar diferentes ficheros de datos en un mismo formato. Consolidar los conocimientos de Python básicos, así como el tratamiento de ficheros en diversos formatos y extracción de datos de una base de datos.
2. Realizar un análisis exploratorio de datos exhaustivo, con el fin de entender los datos de entrada y ser capaces de identificar que datos necesitan ser limpiados mediante técnicas de visualizacion.
3. Implementar *Scrum* como marco de referencia para el desarrollo del producto, basándonos siempre en los valores de *Agile* como puntos clave del trabajo en equipo y la mejora continua.
4. Mejorar la comunicación entre los miembros del equipo.
5. Mejorar vuestras habilidades de comunicación en público al exponer el proyecto en la sesión final.

Caso de uso

Con este proyecto vais a demostrar que sois capaces de unificar un conjunto de datos bajo un mismo archivo, realizar un analisis explotorio de datos y visualizaciones explicativas. Esto os permitirá mostrar vuestras habilidades a enfrentarnos a un conjunto de datos haciendo uso de las librerías más comunes para el tratamiento de datos en Python, algo que os será util a la hora de demostrar vuestros conocimientos a las empresas durante los futuros procesos de selección a los que os enfrentais.

Especificaciones

En desarrollo del procesado del conjunto de datos haremos uso de las siguientes tecnologías y paquetes de Python:

- **Pandas**: Unión de fuentes de datos, procesado y limpieza de datos.
- **Matplotlib/Seaborn**: Para la visualización de datos.
- **Sidetable**: Para obtener estadísticas de los conjuntos de datos.
- **Sckit-Learn**: Para codificar los datos y poder normalizarlos.

El proyecto deberá contener los siguientes elementos:

1. Un repositorio con los ficheros utilizados, así como el código empleado para el desarrollo del proyecto. Siguendo una estructura de carpetas coherente y sencilla.
2. Habrá que asegurarse que todos los archivos finales sean entregados.
3. Elaboración de una presentación para el día de la demo.

Planificación del proyecto

Sprints

Para la realización de este proyecto trabajaremos en 3 *sprints* (3 iteraciones) de entre 5-7 sesiones cada iteración. Siguiendo los principios ágiles, estableceremos pequeños ciclos iterativos de forma que al final de cada uno generemos valor perceptible por nuestros usuarios. Dedicaremos el primer día a la planificación del *sprint* (**sprint planning**) y el resto a trabajar en el desarrollo del proyecto. Al final de cada *sprint* haremos un *Sprint Review* del proyecto para presentar los resultados conseguidos y recoger *feedback*, al igual que una retrospectiva (retro) para evaluar cómo ha ido el *sprint*, además de valorar vuestro trabajo en equipo de cara a mejorar en el siguiente *sprint*.

También haremos una retro corta revisando los *working agreements* que hemos acordado al inicio del proyecto y añadiendo cualquier otro *feedback* que nos permita mejorar el proyecto.

Al final del *sprint* (que coincidirá con el final del proyecto), haremos una sesión de presentación más completa, más allá de lo que sería un *Sprint Review*

Historias de usuario

Para la gestión del proyecto, usaremos historias de usuario, que es una herramienta para definir las características de un producto que veremos en detalle durante el curso.

1. Unión automatizada de los diferentes ficheros de entrada.

Valor

El cliente quiere tener todos los datos en una misma fuente, ya que únicamente va a poder trabajar con un único tipo de datos de ahora en adelante.

Contexto

El cliente ha tenido una reunión con el equipo de datos de la empresa y le han comunicado que debido a problemas internos y mejoras de infraestructura durante el próximo año unicamente van a poder dar soporte a una tipo de datos, por lo tanto van a tener que de forma temporal unificar los diferentes datos disponibles, en lo que mejoran los sistemas.

Criterios de aceptación:

- Crear la infraestructura necesaria: repositorio en GitHub y con acceso para todos los miembros del equipo.
- Unir los diferentes ficheros de entrada en uno que agregue toda la información.
- Unir los diferentes ficheros de entrada en uno que agregue toda la información y guardarlo en csv.
- **DOD: Tener en el repositorio de GitHub el archivo unificado en extension csv.**

2. Limpieza de los datos.

Valor

El cliente quiere tener todos los datos unificados con nombres más descriptivos, hacer una selección de las variables más interesantes y realizar un estudio de las caracteristicas de los datos.

Contexto

En este caso el cliente, no necesita todas las variables que se tienen en el conjunto de datos, ya que se espera poder presentar un informe en el futuro, nos ha pedido que seleccionemos aquellas variables que puedan resultar más interesantes para su análisis.

Criterios de aceptación

- Realizar una selección de las variables que resulten más interesantes de análisis.
- Realizar la limpieza de las columnas seleccionadas.
- Guardar el dataset limpio en un fichero csv diferente al original.
- **DOD: Tener otro conjunto de datos con las columnas que posteriormente realizaremos en análisis en formato csv y en el repositorio de GitHub.**

3. Análisis exploratorio de datos

Valor

El cliente quiere saber que posible información contiene las variables seleccionadas, ya que desea entender en profundidad las peculiaridades de los datos disponibles.

Contexto

Se desea conocer en mayor detalle sobre los datos seleccionados para el análisis, por tanto, le gustaría que se realizase un análisis exploratorio de datos con el fin de identificar datos extraños que deban ser limpiados, antes de proceder a la creación de gráficas explicativas.

Criterios de aceptación

- Obtener algunas estadísticas e información sobre la naturaleza de los datos y guardarlos en un csv o excel.
- Explicar los resultados.
- **DOD: Extraer los datos de las estadísiticas y guardar en fichero externo, las soluciones en el repositorio de GitHub.**

4. Obtención de gráficos

Valor

El cliente quiere obtener conocimiento de la información contenida en los datos que han sido seleccionados para el análisis, para poder dar explicaciones relevantes sobre los datos disponibles.

Contexto

Se desea conocer en mayor detalle sobre los datos seleccionados para el análisis, debido a esto y con vista a la presentación de un informe nos ha pedido que realizemos algunas gráficas con las que posteriormente se pueda explicar algunos detalles relavantes sobre los datos disponibles.

Criterios de aceptación

- Tener como mínimo tres gráficas ilustrativas, se recomienda algún histograma.
- **DOD: Gráficas guardadas en algún ficheros de tipo png,jpeg o jpg. Así como que esten subidas al repositorio de GitHub.**

Ficheros

Los ficheros a descargar son los siguentes:

exported_db_data.csv 1MB
Binary

Descarga este fichero.

data_txt_clean.txt 14MB
Binary

Descarga este fichero.

data_xml_clean.xml 3MB
Binary

Descarga este fichero.

data_remaning_kaggle.csv 16MB
Binary

Descarga este fichero.

kaggle_survey_2021_answer_choices.pdf 158KB
PDF

Descarga este fichero. Para entender las columnas y sus respuestas.

Entrega

El formato de entrega de este proyecto será mediante la subida de este a la plataforma de GitHub. Para subirlo, se creará un repositorio en la organización de Adalab. El nombre del repositorio deberá estar compuesto de las siguientes partes, todo ello separado por guiones:

- La palabra project-da.
- Letra de la promoción promo-B.
- Número del módulo module-2.
- Número del equipo team-X. Por ejemplo:
- Adalab/project-da-promo-b-module-2-team-1
- Adalab/project-da-promo-b-module-2-team-3

En lo relacionado en las fechas de los *sprints*:

- Entrega del primer *sprint* (*sprint* review): 16 Diciembre.
- Entrega del segundo *sprint* : 3 Enero

- Entrega del tercer *sprint* (sprint review): 13 Enero.
- Demo del proyecto (presentación final): 17 Enero.

En las *sprint review* se revisará que se hayan solucionado todas las tareas técnicas asociadas a la entrega de esos *sprints*, si algo quedara pendiente se arrastraría al siguiente *sprint*.

Presentación

El último día del módulo presentaréis la versión final de este proyecto a vuestras compañeras y al equipo de Adalab. Cada equipo realizará una presentación de 5 minutos y posteriormente habrá 5 minutos de *feedback* por parte del público. En este caso, la audiencia podría ser más variada pues no sólo estarán los profesores.

El objetivo es que practiquéis la realización de las demos de los proyectos que habéis desarrollado, explicándolo desde un punto de vista técnica y también desde la perspectiva del producto, mejorando además vuestras habilidades de exposición, objetivo de desarrollo profesional del curso.

Para que la presentación salga bien es imprescindible una buena preparación. Por ello, durante el segundo *sprint* del módulo tendréis que asignar responsabilidades dentro del equipo relacionadas con la preparación de ésta. A continuación incluimos algunos elementos que os pueden ayudar a enfocar la presentación:

- En el público habrá personas con conocimientos técnicos y no técnicos.
- La parte central de la presentación será mostrar el software desarrollado funcionando, a ser posible en directo de forma dinámica o a través de un vídeo (si no fuera posible, como plan B).
- En este módulo, de los diferentes elementos adicionales que os proponemos, sería útil incluir una breve presentación de los diferentes integrantes del equipo desde un punto de vista profesional. Se trata de practicar vuestro "relato" profesional en versión muy corta. Que las personas asistentes conozcan quienes sois como profesionales. Os será también útil para las entrevistas de trabajo.
- Todas las participantes del equipo deben hablar en la presentación (sin práctica no hay mejora).

Además de esto, para mejorar vuestras habilidades de exposición en público y hacer la presentación más rica, podréis incorporar otros elementos adicionales (son solo ideas, sentíos libres de innovar y ser creativas):

- Dejar muy claro quién ha sido vuestro cliente y qué fue lo que os pidió.
- Explicar qué necesidades cubre o qué problemas soluciona el producto, cuál es el beneficio principal que aporta y qué lo hace único comparado con otros productos parecidos del mercado.
- Aportaciones "únicas y diferenciadoras" de cada equipo al proyecto.
- Cómo ha sido la organización del equipo, el reparto de tareas y la coordinación a la hora de trabajar todas en el mismo código.
- Cuál de las tareas o los puntos ha sido el que más esfuerzo ha requerido.
- Cuál de las tareas o partes del proyecto es la que hace que el equipo esté más orgulloso.
- Las tecnologías qué habéis utilizado y para qué sirven, y algunas partes del código que habéis desarrollado que merezca la pena resaltar.

- La presentación debe tener un "buen inicio y un buen cierre" que nos haga a todos estar atentos y aplaudir... ahí os dejamos que echéis a volar vuestra imaginación.
- No habléis en primera persona de lo que habéis hecho, hablad del equipo.
- No mencionéis problemas, sino "retos" que os han hecho aprender y crecer.