1 Practica 1: Recopilación, estructuración y análisis de datos

El objetivo de esta primera practica es el de profundizar en algunos de los conceptos vistos en las sesiones de teoría. La situación que se plantea es la siguiente: acabas de entrar en una gran compañía que gestiona múltiples servicios web. Tu labor dentro de la compañía será analizar algunos procesos del modelo de negocio y desarrollar un pequeño sistema de información. En la empresa tienen definidos una gran variedad de procesos, pero para esta práctica nos interesa el proceso mediante el cual se evalúa debemos ofrecer formación a nuestros empleados sobre seguridad informática y si alguno de ellos supone un riesgo para la compañía.

Nuestra empresa tiene los siguientes departamentos entre otros:

- **Departamento de Recursos Humanos.** La función principal es atraer, desarrollar y retener talento, asegurando un ambiente laboral positivo y el cumplimiento de normativas laborales.
- **Departamento de Analítica Digital.** Utiliza herramientas analíticas para interpretar datos y proporcionar información valiosa que respalde la toma de decisiones estratégicas.
- **Departamento de Ciberseguridad**. Protege la infraestructura tecnológica de la empresa contra amenazas y ataques cibernéticos, garantizando la seguridad de la información.

Comienza cuando el director de Recursos Humanos de la empresa necesita identificar la necesidad de formación sobre seguridad informática de los empleados de cada una de las sucursales de la empresa. Para ello encarga al departamento de Ciberseguridad una campaña de recogida de información sobre los empleados de una sucursal. El departamento monitoriza la actividad de los empleados y analiza los datos de acceso, gestión de contraseñas de los empleados y de registros de correos de phishing detectados. Estos archivos son

1) informe de usuarios con el log de sesiones, hash de sus contraseñas, permisos, datos personales, etc.
2) el listado de las páginas web más visitadas por los empleados de la empresa.

Elabora un informe inicial con posibles problemas de seguridad entre los empleados. Si los datos no son relevantes solicitará la confirmación del director de Recursos Humanos para hacer una campaña interna adicional de phishing para obtener más información de análisis de los empleados, el director de RRHH determinará si es necesario realizar la campaña. En caso negativo se remitirá el informe inicial generado a RRHH y este finalizará el proceso. En caso afirmativo se ordenará dicha campaña al departamento de Ciberseguridad, se adjuntarán los datos a los anteriormente obtenidos y se remitirán al departamento de RRHH.

Una vez recopilada por RRHH toda la información se enviará al departamento de Analítica Digital.

El servicio de Analítica Digital que se encargará de analizar los datos y determinará para que tipo de proceso informático es necesario reforzar la formación en seguridad informática. El departamento de Analítica Digital rastreará las webs más visitadas relacionadas con la formación en seguridad informática de los empleados, identificando patrones y áreas de interés. Se analizarán los datos obtenidos y si fuera necesaria alguna información adicional se contactaría con el departamento de Recursos Humanos para obtenerla. Una vez realizado todo el procesamiento y análisis de datos se enviará un informe final a Recursos Humanos.

Recursos Humanos enviará un email al director de la sucursal identificando el tipo de curso a realizar por los empleados

Si el director de la sucursal da su visto bueno, solicitará dicho curso a RRHH y se matriculará a dichos empleados.
La primera tarea dentro de esta compañía será la de modelar, usando BPMN y UML, este proceso de negocio que se acaba de describir. La segunda tarea que te han encomendado es la de desarrollar, una primera versión de un sistema de información gerencial o MIS (Management Information System). Para ello, se te proporcionan los archivos JSON extraídos de sus bases de datos. Tu labor será la de desarrollar, usando el lenguaje de programación Python y una base de datos SQLite, este pequeño MIS que genere la información que se detalla en los diferentes ejercicios de este enunciado. Además, se pondrá en práctica el uso de librerías pensadas para realizar un análisis de datos que permita generar información interesante para un potencial usuario final, cliente o, simplemente, alguien interesado en recibir la información que se obtiene tras el filtrado y análisis de los datos proporcionados. Concretamente, utilizaremos la librería Pandas para ayudarnos a realizar esta tarea. Es importante que el alumnado realice esta primera practica con éxito y de forma adecuada, ya que la segunda practica se basará en gran medida en los resultados, o avances alcanzados en esta práctica. La práctica se realizará en **grupos de 3 personas**. La práctica solo debe ser entregada por un integrante.

2 Conjunto de datos

Para esta práctica, utilizaremos el archivo de material adicional, el cual contiene diferentes archivos JSON.

- **legal.json** contiene el análisis realizado por el departamento de legal de cada una de las webs.´
- **users.json** contiene la información de cada uno de los usuarios de la empresa.

3 Ejercicio 1 [3 puntos]

En este primer ejercicio, el grupo deberá desarrollar el modelado del proceso de negocio descrito anteriormente usado las dos notaciones vistas en teoría: Business Process Modeling Notation (1.5 punto) y Unified Modeling Language (1.5 puntos)

4 Ejercicio 2 [2 puntos]

El objetivo de este ejercicio será el de desarrollar un sencillo sistema ETL. No es necesario desarrollar las fases de extracción ya que disponemos de los archivos JSON. Debemos diseñar las tablas en la base de datos y desarrollar los códigos necesarios para leer los datos del fichero JSON y almacenarlos en la base de datos. Después, será necesario leer los datos desde la BBDD (usando diferentes consultas) y se almacenaran los resultados en un DataFrame para poder manipularlos. En este ejercicio, para el correcto desarrollo del sistema MIS, será necesario calcular los siguientes valores:

- Numero de muestras (valores distintos de missing).
- Media y desviación estándar del total de fechas en las que se ha cambiado la contraseña.
- Media y desviación estándar del total de IPs que se han detectado.
- Media y desviación estándar del número de email recibidos de phishing en los que ha interactuado cualquier usuario.
- Valor mínimo y valor máximo del total de emails recibidos.

- Valor mínimo y valor máximo del número de emails de phishing en los que ha interactuado un administrador.

5 Ejercicio 3 [2.5 puntos]

Hay datos que nos interesa analizar basándonos en agrupaciones, para darle un sentido a nuestro análisis en base a esa agrupación. De una manera más específica, vamos a trabajar con las siguientes agrupaciones:

- Por tipo de permisos de usuario (0 equivalente a usuario y 1 equivalente a administrador)
- Contraseña débil: estableceremos dos rangos diferentes, el primero aquellos contenidos que tengan una contraseña débil y el segundo los que no. Contraseña débil será aquella que se puede resolver con el diccionario SmallRockYou

En este caso deberemos calcular la siguiente información para la variable dentro del email de phishing:

- Numero de observaciones
- Numero de valores ausentes (missing)
- Mediana
- Media
- Varianza
- Valores máximo y mínimo

6 Ejercicio 4 [2.5 puntos]

Por último, se programarán las diferentes funciones del MIS. En concreto, se deben generar gráficos sencillos para obtener los siguientes datos:

- Mostrar la media de tiempo entre cambios de contraseña por usuario de usuarios normales frente a las de usuarios administradores
- Mostrar los 10 usuarios más críticos (un usuario crítico es aquel usuario que tiene la contraseña débil y además tiene mayor probabilidad de pulsar en un correo de spam), representadas en un grafico de barras.
- Mostrar las 5 páginas web que contienen más políticas (cookies, protección de datos o aviso´ legal) desactualizadas, representadas en un gráfico de barras según las políticas.
- Mostrar según el año de creación las webs que cumplen todas las políticas de privacidad, frente a las que no cumplen la política de privacidad.

7 GitHub

Sera de uso obligatorio la creación de un repositorio público de GitHub para la realización de las practicas con los miembros del grupo. Es recomendable el uso correcto de GitHub, se valorará negativamente la incorporación de todos los datos en un solo commit.

8 Normas de Entrega

La entrega de la practica consistirá en un archivo comprimido con los siguientes ficheros:

- **Carpeta src** del proyecto de PyCharm.
- Archivo SQLite con la base de datos creada.
- Un solo documento con la memoria completa en formato PDF en la que se muestren los ejercicios resueltos.
- La memoria debe incluir el nombre y apellidos de los integrantes del grupo y el enlace al repositorio de GitHub.
- Cualquier memoria entregada fuera de plazo o de forma corrupta será considerada como un 0.

Ante cualquier duda durante la resolución de la práctica, escribir a carlos.contreras@urjc.es vía mail. En caso de no poderse resolver la duda vía mail, se puede concertar una tutoría, siempre y cuando se concierte en un período de **hasta 48 horas antes** de la fecha de entrega de la práctica.
La fecha límite para entregar esta práctica será el 28 de Marzo a las 23:55 y se realizará por la plataforma Aula Virtual.

9 Peso de la practica

La evaluación de esta práctica supondrá un 40% de la nota de prácticas.