Prueba Técnica de Conocimientos

Perfíl Analítico - Científico de Datos

Muchas gracias por tu interés en participar en la convocatoria para pertenecer a la familia de **Cantilever**.

Buscamos personas comprometidas, que se destaquen por un trabajo en equipo y de calidad, con buena actitud de servicio, compromiso además de mucha responsabilidad en sus actividades del día a día, siempre dando prioridad a las necesidades del cliente. Además, buscamos personas con ganas de adquirir nuevos conocimientos y técnicas de modelado estadístico, bases de datos y con grandes habilidades en programación en lenguajes como: Python, R, SQL, entre otros.

En esta prueba encontrarás 5 problemas, de los cuales solo hay que realizar 4 a tu elección. El propósito es medir tus habilidades para manipular datos en tareas a desarrollar dentro de la empresa. ***Los puntos a evaluar son la forma de manipulación y comunicación de los resultados***. No te preocupes por la precisión, nos interesa más saber tu forma de trabajar y comunicar los resultados. Puedes usar cualquier herramienta con la que te sientas más cómodo/a. Ten en cuenta que ***alguien más leerá tu trabajo***, por lo que hay que ***ser claro y tener una buena documentación.***

1. Carga de información y manipulación de los datos.

En este problema se te proporcionará algunos datasets que deberás cargar y manipular como se te indique.

- Carga el archivo excel *Datos Maestros*, únicamente la hoja *Master Data Oficial* y las columnas:
  - Nombre visible Agente
  - AGENTE (OFEI)
  - CENTRAL (dDEC, dSEGDES, dPRU…)
  - Tipo de central (Hidro, Termo, Filo, Menor)
- Obtener los registros que pertenecen al agente EMGESA o EMGESA S.A. y a su vez, Tipo de Central sea 'H' o 'T'.
- Cargar el archivo *d1204.txt* por central (este archivo contiene el nombre de la central y el número de horas trabajadas).
- Realizar un merge de ambos datasets por Central.
- Calcular la suma horizontal de todas las horas de cada planta.
- Seleccionar solo los registros de las plantas cuya suma horizontal sea mayor a cero.

El resultado final debe ser mostrado en algún archivo. De igual forma, el código debe ser presentado explicando paso a paso la solución (ser breve en la explicación).

---

2. Prueba de SQL

SQL es un lenguaje de consultas estructurado. Es de los idiomas más comunes para almacenar, manipular y recuperar bases de datos.

Aquí te proporcionaremos el código para crear algunas tablas y realizar algunas consultas sobre estas.
Te recomendamos usar alguna herramienta de internet:

- https://sqliteonline.com/
- http://www.sqlfiddle.com/

o cualquier otra que te permita usar SQL y sea de tu agrado.
El código siguiente es para la creación de las tablas con las que se debe trabajar (copia, pega y ejecuta el código para la creación de las tablas e inserción de registros).

```
CREATE TABLE EMPLEADO (
  ID INT(8),
  NOMBRE VARCHAR(50),
  APELLIDO VARCHAR(59),
  SEXO CHAR(1),
  FECHA_NACIMIENTO DATE,
  SALARIO DOUBLE(10,2)
  );
```

```
CREATE TABLE VACACIONES(
  ID INT(8),
  ID_EMP INT(8),
  FECHA_INICIO DATE,
  FECHA_FIN DATE,
  ESTADO CHAR(1),
  CANTIDAD_DIAS INT(8)
);

/*EN ESTA TABLA SE ALMACENA LA INFORMACIÓN BASICA DE LOS
EMPLEADOS*/
 INSERT INTO EMPLEADO VALUES (1,"JUAN","PELAEZ","M",'1985-01-
29',3500000);
 INSERT INTO EMPLEADO VALUES
(2,"ANDRES","GARCIA","M",'1975-05-
22',5500000);
 INSERT INTO EMPLEADO VALUES (3,"LAURA","PEREZ","F",'1991-09-
10',2500000);
 INSERT INTO EMPLEADO VALUES
(4,"PEPE","MARTINEZ","M",'1987-12-
01',3800000);
 INSERT INTO EMPLEADO VALUES
(5,"MARGARITA","CORRALES","F",'1990-
07-02',4500000);



/*EN ESTA TABLA SE ALMACENA LAS SOLCITUDES DE VACIONES DE
CADA
EMPLEADO*/
 INSERT INTO VACACIONES  VALUES
(1,1,'2019-07-01','2019-07-15','A',14);
 INSERT INTO VACACIONES  VALUES
(2,2,'2019-03-01','2019-03-15','R',14);
 INSERT INTO VACACIONES  VALUES
(3,2,'2019-04-01','2019-04-15','A',14);
 INSERT INTO VACACIONES  VALUES
(4,2,'2019-08-14','2019-08-20','A',6);
 INSERT INTO VACACIONES  VALUES
(5,3,'2019-08-20','2019-08-25','A',5);
 INSERT INTO VACACIONES  VALUES
(6,3,'2019-12-20','2019-12-31','A',11);
```

Con las tablas creadas anteriormente, obtén las siguientes consultas:

- Seleccione nombre, apellido y salario de todos los empleados.
- Seleccione nombre, apellido y salario de todos los empleados que ganen más de 4 millones.
- Cuente los empleados por sexo.
- Seleccione los empleados que no han hecho solicitud de vacaciones.
- Seleccione los empleados que tengan más de una solicitud de vacaciones y muestre cuantas solicitudes tienen los que cumplen.
- Determine el salario promedio de los empleados.
- Determine la cantidad de días promedio solicitados de vacaciones por cada empleado.
- Seleccione el empleado que mayor cantidad de días de vacaciones ha solicitado, muestre el nombre, apellido y cantidad de días totales solicitados.

Crea un archivo donde guardes los resultados, estos pueden ser presentados con capturas de pantalla y código, copiar el código y captura de la tabla resultante o como más se te acomode. Realiza todas las consultas que puedas, si hay alguna que no lo logras obtener, puedes explicar como la obtendrías sin usar código.

---

3. Prueba de modelación Analítica

Para esta prueba se te dan dos archivos: *train.csv* y *test.csv,* los cuales contienen información sobre transacciones de tarjetas de crédito y débito. En cada transacción se tiene información numérica y cualitativa (revisar *diccionario_variables.csv*).
El archivo *train.csv* contiene los registros etiquetados (si son fraude o no): 1 para FRAUDE y 0 para transacción LEGITIMA.

- El objetivo es desarrollar un modelo (con el archivo *train.csv*) que permita predecir si una transacción es fraudulenta o es legítima, para después evaluarlo con el archivo *test.csv.*

Los resultados de predicción deben ser presentados en un archivo *test_evaluado.csv.*

Nota: utilice el modelo que desee, no importa que sea el más sencillo. Nuevamente, NO estamos evaluando precisión en los resultados, nos interesa la forma en como trabaja y transmite los resultados. Por lo que deberá agregar una explicación de su modelo (breve) y de los resultados obtenidos.
Finalmente, una pequeña sugerencia para mejorar su modelo o qué cambios realizaría.

---

4. Series de tiempo

En este problema deberás cargar el documento *serie_tiempo.csv* donde se presentan valores (*variable1* y *variable2*) por fecha. El objetivo es realizar una predicción de las series de tiempo. A continuación tienes unos pasos que puedes seguir para realizar esta tarea:

- Análisis exploratorio
    - Tendencia (pendiente de los datos)
    - Estacionalidad (Variaciones periódicas)
    - Descripción de las estacionalidades
- Preprocesamiento de los datos.
    - Limpieza y rellenado
    - Transformaciones
- Extracción de características
    - Crea o transforma variables
    - Eliminar variables (si lo ves necesario)
- Propuesta de modelo para predicción
- Evaluación del sistema

También puedes agregar todo aquel análisis extra que consideres relevante.

Pista: para tu modelo de predicción es recomendable que tu serie de tiempo sea estacionaria.

---

5. Problema de salvación

Sabemos que ha sido una prueba entretenida y un poco complicada, pero no queremos que te quedes fuera. Por esa razón se ha agregado este punto extra.
En caso de que no hayas logrado resolver algún problema anterior, o sientas que tu solución no es la más adecuada, puedes investigar algún problema de programación, fenómeno físico o reto matemático donde utilices alguna herramienta o conocimiento que te gustaría destacar y con ello compensar lo anterior. Usa tu creatividad y no olvides explicar bien tu problema.

La familia Cantilever te desea mucho éxito!!!