

Prueba reto

Las unidades de la compañía, tienen como objetivo anual reducir el % de rotación con la finalidad de lograr eficiencias, ahorro y sobre todo en ser el mejor lugar para trabajar. El área de Compensaciones ha propuesto identificar las características de los postulantes que permanezcan en la compañía por más de 6 meses. Las informaciones de las características se encuentran disponibles en los archivos en Excel: Base_1 y Base_2, que contiene algunos datos demográficos y las evaluaciones de las habilidades que se necesitan para una posición de cajero. Cada habilidad tiene la nota (columna Valor), el valor mínimo esperado (columna Esperado), la brecha entre el valor esperado y la nota (columna Brecha), y el porcentaje de cumplimiento (columna Cumplimiento). Las bases también contienen información general sobre el nombre del proceso, nombre del perfil y fechas de creación del proceso o archivo, esta información se encuentra en las filas del 1 al 4 en cada archivo.

Para lograr la propuesta se requiere los siguiente:

Generar un proceso automático en Python que contenga lo siguiente:

1. Sin modificar los archivos Excel, subir automáticamente las bases a dataframe's.
2. Pasar a columnas las filas del 1 al 4 de los archivos en Excel, en los dataframe's del punto 1
3. En los dataframe's, pasar a cabeceras o nombres de columnas, la fila 6 de los archivos en Excel.
4. Para el caso donde existen celdas combinadas modificar los contenidos en los dataframe's de la siguiente manera:

- Celdas combinadas originales del Excel (fila 5 y 6 de los archivos Excel):

Última Experiencia				
Empresa	Cargo	Inicio	Fin	Actividades

- Celdas modificadas en los dataframe's:

Última Experiencia Empresa	Última Experiencia Cargo	Última Experiencia_Inicio	Última Experiencia Fin	Última Experiencia Actividades
----------------------------	--------------------------	---------------------------	------------------------	--------------------------------

5. A cada dataframe crearle las columnas de auditoría similar a estas:

Ejemplo para la Base_1:

RECORD_SOURCE	LOAD_DATE	PROCESS_DATE	CREATION_USER
Base_1.xlsx	2/08/2023 11:37	2/08/2023 13:06	PM999999999
Base_1.xlsx	2/08/2023 11:37	2/08/2023 13:06	PM999999999

RECORD_SOURCE: Nombre del archivo y su extensión.

LOAD_DATE: Fecha y hora de inserción a Python.

PROCESS_DATA: Fecha y hora que culmina de consolidarse las bases.

CREATION_USER: Usuario Windows del ejecutor del script de Python.

6. Consolidar los dataframe's en uno solo.
7. Eliminar los registros que no contengan datos en las competencias: Calidad del trabajo, Desarrollo de relaciones, Escrupulosidad/Minuciosidad, Flexibilidad y Adaptabilidad, Orden y la calidad, Orientación al Logro, Pensamiento Analítico, Resolución de problemas, Tesón y disciplina y Trabajo en equipo.
8. Reemplazar los valores null o vacíos de la columna "Fecha de Finalización de Proceso (Zona horaria GMT 0)" por los valores de la columna "Fecha de Ingreso a Proceso (Zona horaria GMT 0)".
9. Quitar los duplicados usando la columna "No. Identificación" como llave principal. El criterio de limpieza es quedarse con el "No. Identificación" con última "Fecha de Finalización de Proceso (Zona horaria GMT 0)".
10. Añadir la columna "Permanencia>6meses" del archivo Base_permanencia.xlsx al dataframe consolidado.
11. El resultado de la data consolidada debe ser como la base de nombre: "Modelo_base_consolidado.xlsx".

Finalmente, usando los datos de la base consolidada generar un informe en PBI mostrando información relevante respecto a la variable objetivo "Permanencia>6meses".

Opcional, realizar el modelamiento en Python para identificar el mejor algoritmo de machine learning para predecir la permanencia mayor a 6 meses y fundamentar qué criterios usaste para definir por qué elegiste el algoritmo.

Entregables:

- Script de Python, detallando en cada paso, qué acción está realizando.
- El PBI, con la conexión a la base consolidada.
- En caso haya realizado el modelamiento, el script de Python detallando el paso a paso.