ENTREGABLES: Entregar vía correo electrónico, el día estipulado como fecha de entrega, lo siguiente:

1. Archivo que contenga las capturas de pantalla mostrando la evidencia de realización de todas y cada una de las actividades, así como de sus correspondientes resultados.

NOTA: Los archivos deben estar adjuntos al correo electrónico, NO deben ser parte del texto del mensaje.
Objetivo: Profundizar en los métodos de comparación, indexación, clasificación y evaluación de la clasificación con datos etiquetados y la misma estructura.
Contexto: FEBRL en su directorio: \recordlinkage\datasets\febrl presenta varios archivos en donde se tiene controlado el número de registros originales y duplicados. Entiéndase por duplicados dos registros cuando su porcentaje de similitud es por arriba del 50% (dependiendo de los umbrales dentro de la clasificación).

Archivo	Características
dataset1	Este conjunto de datos contiene 1000 registros (500 originales y 500 duplicados, con exactamente un duplicado por registro original
dataset2	Este conjunto de datos contiene 5000 registros (4000 originales y 1000 duplicados), con un máximo de 5 duplicados basados en un registro original (y una distribución de poisson de registros duplicados). Distribución de duplicados: 19 registros originales tienen 5 registros duplicados 47 registros originales tienen 4 registros duplicados 107 registros originales tienen 3 registros duplicados 141 registros originales tienen 2 registros duplicados 114 registros originales tienen 1 registro duplicado 572 registros originales no tienen registro duplicado
dataset3	Este conjunto de datos contiene 5000 registros (2000 originales y 3000 duplicados), con un máximo de 5 duplicados basados en un registro original (y una distribución Zipf de duplicado registros). Distribución de duplicados: 168 registros originales tienen 5 registros duplicados 161 registros originales tienen 4 registros duplicados 212 registros originales tienen 3 registros duplicados 256 registros originales tienen 2 registros duplicados 368 registros originales tienen 1 registro duplicado 1835 registros originales de no tienen registro duplicado

Cada archivo contiene los siguientes campos:

rec_id	given_name	surname	street_number	address_1	address_2	suburb	postcode	state	date_of_birth	soc_sec_id
rec-101-dup-0	amberr	coulson	26		meilene retirement	whitfield	3986	nsw	19950320	3079240
rec-101-org	amber	whillas	3		meilene retirement	whitfield	3987	nsw	19950320	3079240

Como se puede observar, el campo rec_id indica el numero de registro y si es original o duplicado.

Actividad 1: Utilice diversas combinaciones (mínimo 6) de los siguientes métodos hasta que encuentre el mejor modelo para encontrar los registros duplicados en dataset1 (de aquí sacar muestra de entrenamiento y de prueba). Puede utilizar los campos indicados, variarlos o mantenerlos fijos.

Indexado	BlockIndex SortedNeobourhood Q-gram Index	Given_name, surname, suburb, postcode, date_of_birth, soc_sec_id
Comparación	Jaro Levenshtein Q-gram (dice,jaccard)	Given_name, surname, suburb,postcode,date_of_birth,soc_sec_id
Clasificación	FellegiSunter Kmeans Naive Bayes Logistic Regression	
Evaluación de la clasificación	Confusión matrix Precisión Recall Precisión Accuracy specificity	Recuerde que no todos los métodos de evaluación son realmente representativos, debe considerar la distribución de clase, si la ocurrencia de falsos negativos no es aceptable, etc.

Una vez encontrado el mejor modelo, justifica el método de evaluación y presenta su resultado.
Debes presentar los resultados de todas las combinaciones.

Actividad 2: Utilice las diversas combinaciones o el mejor modelo del paso anterior (utilizando la totalidad de dataset1 como entrenamiento) para encontrar los registros duplicados en dataset2.
¿Los resultados fueron mejores o iguales que en el caso anterior? Si—¿por que?, No. ¿por que?