**Formación en modelos de clasificación y comparativas estadísticas**

En este caso crearás un sistema de clasificación para detectar incongruencias en los resultados de análisis de prevención de riesgos laborales de una serie de empresas.

**Enunciado**

En una evaluación de riesgos laborales de una empresa, se desplaza un técnico para realizar el análisis de riesgos y luego redacta una memoria donde indica qué riesgos ha detectado. En la teoría, entre empresas del mismo sector, las evaluaciones de riesgos deberían ser idénticas; pero en la práctica, no es así, ya que los técnicos muchas veces hacen la evaluación "de memoria" y omiten anotar algunos riesgos.

En resumen, lo que ocurre actualmente es que las evaluaciones por sector deberían ser iguales o muy parecidas... pero en la realidad difieren bastante y esto es peligroso, ya que no se están evaluando bien los riesgos.

Así pues, crearemos un sistema en Sk-Learn que, en función del sector de cada empresa (campo nSectorEmp), pueda identificar si una evaluación de riesgos es correcta o no, en base al histórico.

Además del % de similtud, tambíen deberemos identificar qué riesgos suelen aparecer por cada nSectorEmp, y también cuales aparecen en una evaluación de una empresa concreta PERO no acostumban a ocurrir.

Te comparto el dataset "riesgos dataset.xlsx" el cual contiene, para una serie de empresas, una serie de campos de los cuales sólamente debes fijarte en:

* sCodCliente: código interno que se le da a cada empresa (ignora el "idEmpresa", no lo usaremos)
* nSectorEmp: código del sector a que pertenece la empresa
* idEvaluacion: código de cada vez que se le pasa la "evaluación de riesgos laborales" a cada empresa, no es único sino que cada empresa pasa varias evaluaciones y se las enumera correlativamente
* ixRiesgo: riesgo identificado al pasar una evaluación de una empresa
* ixPuesto: puesto donde han identificado el riesgo (p.ej. personal de administración, o bien un fontanero, o alguien del dpt. de logística) - el puesto va vinculado al riesgo
* ixFuenteRL: similar al ixPuesto, pero va asociado a una particularidad de cada situación.

**OJO:** un riesgo o bien va asociado a un puesto, o a una fuente. Nunca irá vinculado a dos elementos, ya verás que si ixPuesto es cero, entonces ixFuenterl no es cero.. y vice versa.

También adjunto un diagrama UML de la BBDD que corresponde al dataset para que puedas entender mejor los datos que hay allí.

Acceso al dataset: https://apioverstand.es/training/riesgos\_ejercicio.zip

¡Ánimos!

**Entregable 1**

Análisis de tecnologías existentes y, si llevan coste asociado, pantallazo de dicho coste. Se dará prioridad a las tecnologías open source que no lleven costes asociados.

En base a ello, se decidirá por la tecnología a usar y se procederá al desarrollo de la solución.

**Entregable 2**

Entregables (en un único wetransfer):

1. Código fuente de la solución
2. Video explicativo del código fuente (máximo 3 minutos)
3. Video mostrando e interpretando los resultados (máximo 2 minutos)

¡Ánimos!