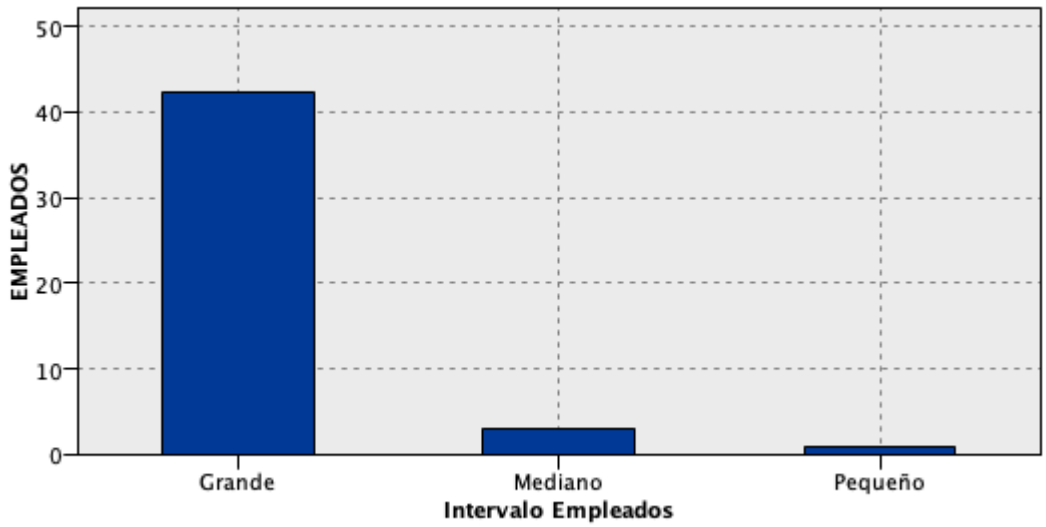
PREPARACIÓN DE DATOS

Para la parte de preparación de datos, trabajaremos con el fichero Dataset de Iberinform, este cuenta con registros de 21.562 empresas españolas en el que se les otorga un scoring que tiene como objetivo realizar un análisis de riesgo crediticio.

En primer lugar, hemos creado la variable objetivo: Riesgo. Esta variable procede de la columna score que contiene valores del 0 al 9. Tal y como indica el enunciado hemos determinado que los valores mayores a 4 no serán empresas de riesgo, por lo que atribuimos un valor Falso, en caso contrario nos encontramos las empresas que tienen un score menor de cuatro que serán denominadas empresas con riesgo crediticio, por lo que les hemos otorgamos el valor True. Tras crear esta nueva variable, le modificamos el tipo a marca y eliminamos la columna score, puesto que su contenido no será necesario en el modelo.

En segundo lugar, hemos trabajado sobre la variable empleados, ya que es relevante a la hora de dictaminar el tamaño de la empresa. Lo primero que hemos hecho es dividir esta variable en tres grupos a través de los terciles y atribuyendo a cada grupo, una etiqueta de pequeña, mediana o grande empresa.



En tercer lugar, hemos trabajado sobre la variable concurso de acreedores para eliminar los valores nulls que contenía. Para ello, hemos reclasificado los valores nulos por “no” y hemos mantenido los valores “si”.
En cuarto lugar, para obtener información más detallada sobre la situación financiera de las empresas, hemos creado la columna “Pasivo”. Esta se basa, en la resta de Activo - Patrimonio Neto.

En quinto lugar, hemos unido las variables: “Número de reclamaciones” y “Número de incidencias” en una nueva columna llamada incidencias totales. Previo a la creación de esta columna, hemos modificado los valores blancos que contenían las dos variables y los hemos sustituido por “0" para poder realizar la operación.

Una vez realizadas estas modificaciones hemos procedido a filtrar algunas variables, y hemos seleccionado las 14 siguientes, todas ellas tienen rol de entrada menos la variable riesgo que tiene rol de destino.

- Forma Jurídica
- Provincia
- Ventas
- Resultados
- Activo
- Patrimonio Neto
- Capital Social
- Endeudamiento a corto plazo
- Fondo maniobra
- Riesgo
- Intervalo empleados
- Concurso
- Pasivo
- Incidencias totales

Una vez realizado este filtro manual, hemos realizado un filtro automático, para verificar que las variables seleccionadas son relevantes. Con este filtro, nos hemos percatado que las variables ventas y resultados no eran necesarias para este modelo, por lo que, en consecuencia, las hemos eliminado.

Para finalizar con la parte de preparación de datos, hemos dividido la base de datos en entrenamiento, comprobación y validación, dando un porcentaje de 60%,30%,10% respectivamente.