



UNIVERSIDAD DE CASTILLA-LA MANCHA

ESCUELA SUPERIOR DE INGENIERÍA INFORMÁTICA

GRADO EN INGENIERÍA INFORMÁTICA

**TECNOLOGÍA ESPECÍFICA DE
COMPUTACIÓN**

TRABAJO FIN DE GRADO

**Predicción del fracaso empresarial mediante ciencia de
datos**

Jaime Tolosa De La Fuente

Agosto de 2019





UNIVERSIDAD DE CASTILLA-LA MANCHA

ESCUELA SUPERIOR DE INGENIERÍA INFORMÁTICA

GRADO EN INGENIERÍA INFORMÁTICA

**TECNOLOGÍA ESPECÍFICA DE
COMPUTACIÓN**

TRABAJO FIN DE GRADO

**Predicción del fracaso empresarial mediante ciencia de
datos**

Autor: Jaime Tolosa De La Fuente

Directores: José Antonio Gámez Martín

Agosto de 2019

A mi familia.

Declaración de Autoría

Yo, Jaime Tolosa De La Fuente con DNI 49313350Q , declaro que soy el único autor del trabajo fin de grado titulado Predicción del fracaso empresarial mediante ciencia de datos y que el citado trabajo no infringe las leyes en vigor sobre propiedad intelectual y que todo el material no original contenido en dicho trabajo está apropiadamente atribuido a sus legítimos autores.

Albacete, a 11 de Junio de 2019

Fdo.: Jaime Tolosa De La Fuente

Resumen

El estudio del fracaso empresarial ha sido objeto de estudio desde los años sesenta, ya que nos encontramos en un mercado económico que es muy poco previsible.

El riesgo empresarial puede ser medible por ratios financieros. En los tiempos que corren, este tema ha sido de gran interés tanto para empresarios, la banca y muchas otras entidades. Por ello diversos estadísticos y economistas han intentando mejorar los modelos ya existentes con el fin de poder predecir la quiebra de una empresa y subsanarla antes de que esta ocurra, además de poder observar que causas están provocandola. Con el objetivo de que esto se pueda predecir de manera automática surge la idea de este Trabajo Fin Grado, donde se realizará un estudio de ciencia de datos, sobre una base de datos llamada SABI disponible para la UCLM, donde identificaremos los procesos y algoritmos para poder calcular la probabilidad del fracaso empresarial de una empresa.

Abstract

Business failure study's have been studied since the sixties. Now, exist an economic marketwhich is very unpredictable.

Bussiness risk can be measured by financial ratios. Now, this topic is very interesting to bussinesman, banking and others entities. Accordingly, a lot of stadistics and economists have tried improve the models that already exists. They wanted to predict the failure bussiness and correct it before it happens. With this objective emerges this project, where it will be done a study of data science, about a data base called SABI available to UCLM. This data base will identify the processes and algorithms to calculate the bussiness failure's probability of a company

Agradecimientos

Con este Trabajo Fin de Grado, finaliza una de las etapas en las que más me he realizado como persona. En primer lugar, querría agradecer esto a mi familia y a mi pareja. Ellos han hecho que a día de hoy este donde estoy, pero sobre todo os agradezco que me hayáis apoyado y respetado todas las decisiones que he tomado.

También quiero agradecer el apoyo de mis amigos aquí en la facultad. Gracias por haber sido un grupo de amigos y no de compañeros, habéis hecho todo el camino hasta aquí más fácil.

Por último, quiero agradecer la realización de este Trabajo Fin de Grado a mi tutor José Antonio Gámez, por haber atendido mis dudas en el momento que fuera, pero sobre todo por haber sido uno de los profesores que me motivo ha introducirme en el mundo de la minería de datos y la inteligencia artificial.

Índice general

ÍNDICE DE FIGURAS	xiii
Lista de Figuras	xv
ÍNDICE DE TABLAS	xv
Lista de Tablas	1
1. INTRODUCCIÓN	1
1.1. Motivación	2
1.2. Objetivos	4
1.3. Estructura de la memoria	5
2. ANTECEDENTES Y ESTADO DE LA CUESTIÓN	7
2.1. Fracaso empresarial y estudios relacionados.	8
2.2. Hot topics en Aprendizaje Automático.	18
3. METODOLOGÍA Y DESARROLLO	33
3.1. Descripción de la metodología	34
4. Fases de trabajo	37
4.1. Primera iteración	38
4.2. Segunda iteración	54
4.3. Tercera iteración	56
4.4. Cuarta iteración.	74
5. Comparación de modelos y el conocimiento del experto	95
5.1. Árboles de Decisión	96
5.2. CN2. Algoritmo de inducción de reglas.	99
6. CONCLUSIONES Y PROPUESTAS	103
6.1. Conclusiones	103
6.2. Trabajo futuro	105
BIBLIOGRAFIA	109

CONTENIDO DEL CD	111
A. EJEMPLO DE USO DE LA DE LA BASE DE DATOS SABI	113
B. COMPETENCIAS DE LA TECNOLOGÍA CURSADA.	117

ÍNDICE DE FIGURAS

2.1. Procedimiento del análisis univariado	14
2.2. Regresión Lineal para clasificación	15
2.3. Regresión Logística para clasificación	16
2.4. Modelo Naive Bayes	17
2.5. Árbol de decisión	18
2.6. Inteligencia artificial conjunto	20
2.7. Perceptron propuesto por Rosenblatt	21
2.8. Problema de clasificación resuelto de manera tradicional.	23
2.9. Problema de clasificación resuelto con un red neuronal.	24
2.10. Red neuronal recurrente básica.	26
2.11. Red neuronal recurrente básica desenrollada.	26
2.12. Red neuronal recurrente interna básica desenrollada.	27
2.13. LSTM desenrollada.	27
2.14. Efecto del bagging en la clasificación	29
2.15. Efecto del boosting en la clasificación	31
3.1. Proceso completo de minería de datos.	35
4.1. Etiquetas empresas Activas en SABI.	40
4.2. Etiquetas empresas Inactivas en SABI.	40
4.3. Evolución del PIB en España.	42
4.4. Evolución del porcentaje de parados en España.	42
4.5. Evolución del número de empresas en España.	43
4.6. Pantalla inicial de la base de datos de SABI.	44
4.7. Número de resultados una vez haber ejecutado los filtros en SABI	44
4.8. Empresas clasificadas como activas	45
4.9. Columnas de las empresas	45
4.10. Exportacion del dataset	46
4.11. Fichero de las actividades económicas según el CNAE.	47
4.12. Frecuencia de empresas fracasadas según el tipo de actividad.	48
4.13. Frecuencia de empresas activas según el tipo de actividad.	50

4.14. Visualización del Árbol de Decisión	55
4.15. Ordenación de la base de datos de SABI	56
4.16. Logaritmo del activo total	57
4.17. Distribución del activo total de empresas en concurso	58
4.18. Diagrama de cajas del tamaño de las empresas.	58
4.19. Como hacer filtros en la plataforma SABI.	59
4.20. Filtros de la muestra	61
4.21. Número de empresas según su tipo de actividad.	63
4.22. Número de empresas según su forma jurídica.	64
4.23. Tiempos de algoritmos de selección de variables	68
4.24. Árbol de la cuarta iteración.	72
4.25. Desglose de los resultados.	75
4.26. Filtro de empresas activas en la ultima iteración.	76
4.27. Filtro de empresas en concurso en la ultima iteración.	77
4.28. Evolución del $f_1 score$ respecto de la β	79
4.29. Ejemplo de árbol de operaciones generadas por un algoritmo genético. .	83
4.30. Ejemplo del individuo de un algoritmo genético.	84
4.31. Proceso del conjunto de datos a las series temporales.	86
4.32. Ejemplo de uso LSTM.	88
4.33. Evolución de la función de perdida.	88
4.34. Evolución del acierto.	89
5.1. Árbol de decisión presentado al experto.	98
5.2. Reglas generadas por CN2.	100
A.1. Bases de datos de la UCLM.	113
A.2. Acceso Red Iris.	114
A.3. Pantalla principal SABI.	114
A.4. Conjunto de empresas SABI.	115
A.5. Unidades de las variables.	115
A.6. Cuadro de diálogo de exportación.	115

ÍNDICE DE TABLAS

2.1. Frecuencia del tipo de variables utilizadas en diversos estudios.	12
2.2. Variables usadas en diversos estudios desde 1966 hasta 2009	12
4.1. Actividades por letra CNAE 2009.	49
4.2. Suma de la diferencias de los cuadrados según el modelo de regresión. .	52
4.3. Reducción del número de nulos tras aplicar el imputador.	52
4.4. Datos del diagrama de cajas del tamaño de las empresas	59
4.5. Datos del diagrama de cajas de diferentes variables	60
4.6. Frecuencia de las formas jurídicas.	64
4.7. Resultados de la selección de variables usando como métrica la precisión.	69
4.8. Matriz de confusión.	70
4.9. Resultados de la selección de variables usando como métrica la f_1score	71
4.10. Frecuencia del tipo de variables utilizadas en la última iteración del conjunto de dato SABI.	78
4.11. Número de variables por método de selección.	79
4.12. Resultados de la selección de variables usando como métrica el acierto en la última iteración.	80
4.13. Resultados de la selección de variables usando como métrica el f_1score en la última iteración.	80
4.14. Conjunto de variables final.	81
4.15. Resultados del algoritmo genético utilizando el acierto y el f_1score en una validación cruzada.	84
4.16. Resultados con el conjunto de evaulación	93
B.1. Tabla de competencias.	118

Capítulo 1

INTRODUCCIÓN

En este capítulo se expondrán los motivos y aspiraciones que han llevado a la realización de este Trabajo Final de Grado(TFG). Además se hará una visión amplia en el contexto que estamos desarrollando este trabajo de investigación. Para finalizar se detallarán los objetivos finales de este TFG, además de la explicación de como se estructurará este documento y de que partes constará.

1.1. Motivación

La predicción del fracaso empresarial ha sido uno de los principales problemas en ámbito de las ciencias empresariales, ya que esto puede ser una gran ayuda en la toma de decisiones de una empresa. Desde la década de los 60 con los trabajos de ([Beaver, 1966](#)) y ([Altman, 1968](#)), se ha intentado buscar nuevas metodologías y modelos para mejorar el resultado de la predicción de la continuidad empresarial, pero no solo eso, sino también buscar los motivos por los cuales se da este fracaso empresarial.

Aunque este problema lo podemos abordar desde diferentes perspectivas de las ramas de la ciencia en este caso lo haremos desde la disciplina del aprendizaje automático y la ciencia de datos.

Actualmente vivimos en la era de la Sociedad de la Información, donde se producen grandes volúmenes de datos. Todos estos datos vienen de fuentes como pueden ser las transacciones, sensores, usuarios de una determinada aplicación o incluso nuestro propio genoma humano. Sabiendo todo esto, ahora debemos de conocer qué podemos hacer con todos estos datos, para ello existe la disciplina de minería de datos o ciencia de datos. Esta nos provee de herramientas con las cuales podemos darle forma y significado a estos datos, para así obtener una ventaja competitiva del conocimiento extraído. Esta no es una tarea nada sencilla, pero para ello poseemos herramientas como el aprendizaje automático que es núcleo o la pieza en torno a la que gira este proceso.

El aprendizaje automático consiste en dotar a las computadoras de la capacidad de aprender sin ser programadas explícitamente para ello, mediante el desarrollo de algoritmos que toman como entrada un conjunto de datos, a partir de estos averiguan unos patrones que dan lugar a predicciones que pueden ser útiles o tener un significado para el científico de datos.

Aunque el aprendizaje automático es la pieza clave, alrededor suya, existen otras tareas en este proceso de minería o ciencia de datos como puede ser el preprocesamiento de estos mismos, visualización, análisis, etc. Aunque necesitemos los algoritmos necesarios para poder encontrar estos patrones, otra parte no menos importante en la que se invierte la mayoría del tiempo en la minería de datos es en el preprocesado de los datos, donde es recomendable en la medida de lo posible que cumplan estos requisitos:

- Completos: Que no existan valores perdidos.
- Sin ruido: No deben existir ni errores ni outliers.
- Consistentes: No deben contener discrepancias(códigos, nombres, etc.)

Muchos algoritmos son robustos a este tipo de datos, por lo que una correcta parametrización y selección del modelo, puede hacernos que no sea necesario invertir tanto tiempo en el preprocesado de los datos. Pero si es cierto, que del proceso del preprocesado de los datos depende que los patrones encontrados por los modelos sean de buena o mala calidad.

Lógicamente, uno de los objetivos de este TFG será poder optimizar el acierto global de la clasificación del modelo, pero también sería importante, en particular, disminuir el tipo de error que clasifica a una empresa como sana que próximamente fallará.

En muchas ocasiones modelos con un buen acierto en clasificación, en realidad no son buenos modelos, esto es algo que se observa bien con ejemplos. Imaginemos que el problema a resolver es sobre la predicción de si una persona tiene un tumor benigno o un tumor maligno. En este caso tenemos un clasificador que devuelve siempre como respuesta la clase más frecuente, que clasificaría con un acierto del 95 % que ningún individuo tiene un tumor maligno en un conjunto de datos donde 950 de las personas son personas con tumores benignos y solo 50 son malignos. Pero sin embargo nunca clasifica como enferma a una persona que realmente si lo está, es decir, que tenemos un modelo más que pobre. Por ello, a la hora de construir modelos no solo es importante observar en cuantos casos te equivocas y en cuantos aciertas sino también tener en cuenta en que casos nos estamos equivocando y en cuales estamos acertando.

En cambio, en nuestro problema es mejor disminuir el error que clasifica a una empresa como sana, cuando esta próximamente se va a encontrar en concurso de acreedores. Ya que si clasificamos a una empresa como no sana, cuando si lo es, la mayor pérdida que podemos obtener es no conceder a esa empresa financiación o un crédito. En cambio, clasificar a una empresa como sana cuando lo más probable es que se convierta en una empresa morosa por no poder afrontar el pago de ese crédito, tiene un mayor coste que el anterior caso. Por ello, es bueno tener en cuenta estas dos métricas a la hora de evaluar un modelo y no solo el acierto, para resolver este problema usaremos otro tipo de métricas que no sea el acierto global.

La base de datos con la que trabajaremos, es la denominada SABI (Sistema de Análisis de Balances Ibéricos)¹ que se encuentra disponible para usuarios de la UCLM. En ella encontramos diferentes ratios financieros que serán significativos a la hora de predecir. En este Trabajo Fin de Grado tendremos en cuenta datos como el tamaño y sector al que se dedica esta empresa. Se hará un preprocesamiento de los datos donde

¹<https://sabi.bvdinfo.com/sso.aspx?path=rediris>

se tratará el ruido de los mismos, valores perdidos, datos inconsistentes, etc. Además, no solo se tratará la limpieza de los datos en el preprocesamiento sino también la construcción de nuevas variables a partir de los ya existentes.

Le daremos una especial importancia a la explicación porque los ratios financieros son nuestras principales variables predictoras. Esto es así, porque la mayoría de la literatura escrita en este ámbito usaba estas variables como predictores con unos resultados más que satisfactorios. Esto, es algo más que lógico puesto que los ratios financieros son coeficientes que proporcionan unidades financieras de medida. De hecho, a menudo estos ratios financieros se usan como medio para evaluar el estado financiero global de las empresas, que es el principal indicador de la solvencia de las mismas. Aunque, como hemos mencionado antes se usarán otro tipo de variables como el sector o el tamaño de la empresa que podemos medir con diferentes parámetros asociados a la misma.

Para finalizar, mencionaremos algo muy importante y es que la variable clase, en un principio tiene 4 etiquetas aunque, debido a la naturaleza del problema, 3 de ellas se fusionarán en una misma etiqueta ya que las empresas que se encuentran en concurso de acreedores, suspensión de pagos o quiebra, son empresas evocadas al fracaso empresarial. Con lo cual, al final solo tendremos dos etiquetas en la variable clase activa e inactiva, en la ultima se englobarán las 3 etiquetas anteriormente mencionadas. Para este problema, en el caso de las empresas donde su situación fiscal sea quiebra, suspensión de pagos o concurso de acreedores se cogerán los datos correspondientes al año anterior de declarar uno de los tres estados correspondientes con el fin de predecir aquellas que en un tiempo muy próximo declarará este estado fiscal. Por la otra parte, para las empresas clasificadas como activas, si obtienen beneficios negativos durante 3 años consecutivos anteriores al año que realizamos el estudio, no se cogerán como ejemplo de empresa activa, puesto que al obtener beneficios negativos de una manera constante, en cualquier momento puede ser declarada en suspensión de pagos, quiebra o concurso de acreedores.

1.2. Objetivos

1.2.1. Objetivos

El objetivo principal del trabajo sera encontrar un modelo a partir del refinamiento de los datos fiscales de las empresas aportados por la base de datos SABI con el que poder predecir si una empresa esta próxima a la bancarrota o no.

De este objetivo principal surgen sub-objetivos parciales que ayudarán a cumplir

este objetivo principal:

- Un correcto preprocesamiento y transformación de los datos: Que nos ayudarán a obtener mejores modelos.
- Correcto análisis y visualización de los datos: Una de las tareas más importantes puesto que nos ayudará a identificar los patrones seguidos por el conjunto de datos.
- Selección de técnicas de minería de datos más adecuadas al problema y obtención de modelos predictivos mediante su aplicación.
- Evaluación de los modelos obtenidos: Se llevará a cabo una evaluación de los modelos obtenidos para así seleccionar el que nos garantice la mayor fiabilidad a nuestro sistema.
- Extraer información del resultado final: Obtener conclusiones a partir del resultado final de cuales son las causas provocan el fracaso empresarial, desde el punto de vista del dominio del problema.

1.3. Estructura de la memoria

- **CAPÍTULO 1. Introducción:** Se presentan las motivaciones y objetivos de este TFG. Además de indicar de las partes de las que consta.
- **CAPÍTULO 2. Antecedentes y estado del arte:** Se indican los antecedentes del TFG y el contexto en el que se desarrolla el mismo. Se exponen los conceptos claves en los que se trabajará en este TFG relacionados con la Minería de datos, el Aprendizaje automático y la Inteligencia Artificial. Además, se mostrará diferentes investigaciones, donde se tomarán muchas de sus metodologías como ejemplo y discrepancias con las mismas. Se indicarán los antecedentes del TFG y el contexto en el que se desarrolla el mismo.
- **CAPÍTULO 3. Metodología y desarrollo:** En este capítulo explicaremos que metodología usaremos para afrontar este problema y de que fases consta.
- **CAPÍTULO 4. Solución del problema:** Aquí, se enumerarán las fases de cada iteración del proceso de minería de datos y se explicará que se ha llevado a cabo en cada fase del proceso.
- **CAPÍTULO 5. Comparación de modelos y el conocimiento del experto:** Se compararán las reglas obtenidas por árboles y sistemas basados en reglas,

para presentarselo al experto y comparar el conocimiento de este con el resultado de los mismos.

- **CAPÍTULO 6. Conclusiones y propuestas:** Se exponen las conclusiones obtenidas de la realización de este trabajo y posibles mejoras que se le pueden dar al mismo.

Capítulo 2

ANTECEDENTES Y ESTADO DE LA CUESTIÓN

En esta sección del TFG describiremos los antecedentes, para poder observar como han evolucionado las metodologías a lo largo de los últimos sesenta años. También, describiremos los modelos más utilizados en los antecedentes y nuevos modelos que aplicaremos en este TFG para este tipo de problemas.

2.1. Fracaso empresarial y estudios relacionados.

2.1.1. Contexto y estudios relacionados con el fracaso empresarial.

Todos los estudios realizados desde la década de finales de los 60, toman como estudio base el trabajo de ([Altman, 1968](#)) y ([Beaver, 1966](#)), y por primera vez se propone un estudio multivariable sobre la bancarrota de las empresas, como en este caso se escogen como únicos valores para la variable clase: fracasada y activa. En el caso del estudio de ([Altman, 1968](#)) separa su conjunto de datos en una parte de entrenamiento y otra de test. En el test con dos datos financieros del ejercicio en el que la empresa se declarará en quiebra obtiene un 95 % de acierto, mientras que cuando lo hace con los datos del ejercicio de dos años anteriores a la declaración de la quiebra obtiene un alrededor de un 72 % de acierto.

Aunque obtiene unos resultados más que satisfactorios, tiene ciertas deficiencias en su ejecución como:

- El número bajo de empresas utilizadas en el estudio: La muestras son aproximadamente de 60 empresas.
- El estudio se hace sobre una muestra equilibrada respecto a la clase: Con lo cual no hace una representación real de como es el problema ,además para Altman la falta de liquidez ya clasifica a una empresa como fracasada, lo cual puede hacer una selección de la muestra muy sesgada.
- El modelo se genera a partir de la muestra del año del fracasado de la empresa y no de año anteriores: Con lo cual su modelo en un principio esta preparado para clasificar empresas fracasadas y no fracasadas pero no para predecir empresas que en un tiempo próximo fracasaran. Esto se denota en las pruebas hechas, porque se pasa de un porcentaje del 95 % de acierto a un 72 % siendo la misma muestra de empresas.

Durante las siguiente décadas, los investigadores creyeron que se podrían obtener modelos sólidos mediante técnicas de estadísticas y de minería de datos. Pero como exponen en estudios ([Montaño, 2009](#)), y ([Tascón and Castaño, 2010](#)) exponen que con el transcurso de los años no se consiguieron modelos sólidos, ni teoría, ni un mapa conceptual de las relaciones explicativas del comportamiento económico de como funciona una empresa, donde se pudieran sentar las bases teóricas del fracaso empresarial.

Aunque como veremos en el trabajo de ([Tascón and Castaño, 2010](#)), que hace un recorrido de como ha sido la evolución de los estudios sobre este campo, si hay consenso

en las variables escogidas para la predicción del fracaso empresarial y se ve que hay una convergencia en el uso de los ratios de rentabilidad y ratios de endeudamiento de las empresas.

Aunque parezca evidentemente debemos dar una breve introducción sobre lo que es el fracaso empresarial, para contextualizar en el marco que nos movemos, aunque en posteriores capítulos lo explique con mayor profundidad. En resumen, tenemos tres tipos de fracaso aquel que se da cuando la empresa es incapaz de hacer frente a sus obligaciones financieras, aquellas en las que la empresa, de manera fiscal se declaran en quiebra, suspensión de pagos o como se unifico por la Ley Concursal de 2014 en España en concurso de acreedores, y finalmente como define ([Altman, 1968](#)), aquella que tiene una falta de liquidez de manera constante durante los diferentes ejercicios.

De hecho, definir de que manera el estudio enfoca el concepto de fracaso empresarial, es algo de suma importancia puesto condiciona de que manera y como hacemos la selección de la muestra del estudio, lo que condiciona a su vez de manera drástica los resultados finales de nuestros modelos.

Esto es algo evidente puesto que escoger el estado fiscal de una empresa nos aporta un concepto mucho más flexible que el que nos aporta por ejemplo la liquidez o el riesgo financiero de la misma. Puesto que, podemos escoger en nuestra muestra empresas activas que presentan ciertas dificultades en uno de los dos aspectos mencionados anteriormente y seguir como empresa activa y viceversa. De hecho, en nuestro caso y aunque los resultados de nuestros modelos puedan verse mermados en cuanto a su eficacia, escogeremos el concepto fiscal, puesto que es interesante ver si los propios modelos de aprendizaje automático son capaces de obtener esa información por si solos. De esta manera, no estamos sesgando las muestras que pueden hacer que nos encontremos con que nuestros datos son "muy fáciles", y trabajamos con muestras de entrenamiento más cercanas a la realidad.

Escoger el fracaso fiscal para clasificar una empresa como fracasada o no también tiene inconvenientes. Puesto que esta definición nos esta hablando de un fracaso empresarial, que ya ha ocurrido y es permanente. Con lo cual los modelos generados, si escogemos los datos financieros del año donde la empresa se declara en quiebra, tendrían muy buenos resultados si el problema a predecir fuera el de clasificar empresas que han fracasado, y disminuiría su eficacia si el problema a predecir fuera el de clasificar aquellas que están próximas al fracaso. Esto se puede solucionar, escogiendo los datos financieros de ejercicios anteriores a la declaración de la quiebra de esa empresa.

Otro problema relacionado con la selección de la muestra y que sesga aun más nuestros modelos, y por lo cual no son han sido generalizables es que el problema es un problema totalmente desbalanceado, en cuanto a nuestra variable clase, con lo cual hace que le sea más difícil a los modelos predictivos encontrar los patrones que definen a una empresa como fracasada. Además, del sesgo que se da, como apunta el estudio de ([Tascón and Castaño, 2010](#)), puesto que estamos estudiando empresas que tengan cierta edad, para nuestra selección necesitamos varios ejercicios de la empresa dejamos fuera a empresas jóvenes que han fracasado o están activas, este es otro motivo más por los que los modelos siguen dejando de ser generalizables para toda la población.

Si nos damos cuenta, las definiciones de fracaso empresarial, no son más que la evolución de los síntomas del fracaso empresarial y en que estado lo escogemos. La definición aportada por ([Altman, 1968](#)) sobre la liquidez no es mas que el signo más temprano del fracaso empresarial, que sería cuando la empresa por falta de liquidez es incapaz de hacer frente a su pasivo líquido. Posteriormente, y cuando esta sigue obteniendo una falta de liquidez llegaría el fracaso financiero, porque es incapaz de hacer frente al pasivo líquido y al fijo, es decir, a la deuda total que posee la empresa, y finalmente el fracaso fiscal que es cuando la empresa de manera fiscal se declara en quiebra.

2.1.2. Variables explicativas del fracaso empresarial.

En los estudios realizados por ([Altman, 1968](#)) y ([Beaver, 1966](#)) obtienen una primera lista, bastante reducida en cuanto al número de variables a partir de su juicio como expertos, y posteriormente reducen en un subconjunto, en uno menor en el caso de ([Altman, 1968](#)) midiendo con estadísticos la capacidad explicativa de un subconjunto de variables, y en el caso de ([Beaver, 1966](#)) observando aquellos que tienen una mayor significatividad de manera individual. Durante los años posteriores, se observa como la manera de seleccionar las variables pasar de ser de una manera más purista a partir del razonamiento económico, a hacerlo de una manera más empírica, es decir, coger todo el conjunto de variables posibles y optar por hacer una reducción de la dimensionalidad a partir de métodos estadísticos o por prueba y error, como apunta el estudio de ([Tascón and Castaño, 2010](#)).

Hay diferentes tipos de ratios para medir el estado de una empresa, como hemos mencionado con anterioridad el tipo de ratio, que se ha ido utilizando en este tipo de estudios ha sido los ratios relacionados con la rentabilidad y el endeudamiento de la empresa. Aunque, no deja de ser menos importante otro tipo de ratios como se repre-

senta en la tabla 2.1 . Para consultar que significada cada ratio una buena referencia sería ([areapymes, 2015](#))¹ o ([economipedia, 2019](#))².

Ahora vamos a definir que es cada tipo de ratio y para que se utiliza:

- Rentabilidad: Este tipo de ratio se utiliza para enfrentar resultados de la empresa en distintos balances o de la cuenta de pérdidas y ganancias. En definitiva, sirven para medir el nivel de eficiencia en la utilización de los activos, en relación con la gestión de sus operaciones.
- Endeudamiento: Los ratios de endeudamiento sirven para medir la calidad del endeudamiento de la empresa, es decir, ver en que medida el beneficio obtenido por la empresa es capaz de soportar el peso de la deuda.
- Equilibrio Económico-Financiero: Los ratios de equilibrio económico-financiero de una empresa representan la eficacia que tiene una empresa a la hora de poder hacerle frente al pasivo en el vencimiento fijado, y por tanto ser capaces de tener un ciclo normal de sus operaciones. Estos ratios son aquellos que relacionan las masas de activo con las masas de financiación.
- Estructura económica: Los ratios, de estructura económica relacionan la estructura porcentual del activo y del pasivo, es decir, describe en porcentaje de que manera la estructura de los activos es capaz de hacerle frente a la de los pasivos.
- Rotación: Este tipo de ratios son aquellos que nos indican la eficacia que tiene la empresa del uso de sus activos para generar rendimientos.
- Otros ratios: Son unos tipos de ratios que representan una menor frecuencia que la familia de los anteriores y poder representar la tabla 2.1 de la manera mas homogénea posible.

En la tabla 2.2 vemos las 21 variables más usadas por los diferentes estudios desde 1966 hasta 2009 según el estudio de ([Tascón and Castaño, 2010](#)), y que serán las que nosotros usaremos para ver comparar su valor predictivo frente a las seleccionadas mediante métodos estadísticos y metodos wrapper. Posteriormente y según los resultados obtenidos se determinará cual será nuestro conjunto de datos final.

¹<https://www.areadepymes.com/?tit=ratios-del-balance-y-de-la-cuenta-de-resultados&name=Manuales&fid=eh00015>

²<https://economipedia.com>

Categorías de ratios	Items	%Total
Rentabilidad	64	17.44 %
Endeudamiento	55	14.99 %
Equilibrio Económico-Financiero	42	11.44 %
Estructura Económica	38	10.35 %
Margen	35	9.54 %
Rotaciones	32	8.72 %
Otros ratios	76	20.71 %
Variables	25	6.81 %
Totales	367	100.00 %

Tabla 2.1: Frecuencia del tipo de variables utilizadas en diversos estudios.

Variables explicativas	Numero de trabajos
Deuda total/Activo total	18
Activo circulante/Pasivo circulante	14
BAIT/Activo Total	14
Beneficio Neto/Activo total	14
Activo circulante/Activo Total	10
Gastos financieros/Pasivo exigible	7
Pasivo exigible/Fondos propios	6
Activo circulante/Pasivo circulante	6
Resultado neto/Fondos propios	6
Activo fijo/Pasivo circulante	5
Cash Flow/Deuda Total	5
Ingresos de explotación/Activo fijo	5
Resultado antes de impuestos/Fondos propios	5
Activo fijo/Pasivo fijo	4
Gastos financieros/Ingresos de explotación	4
Cash Flow/Pasivo circulante	4
Cash Flow/Pasivo fijo	4
Ingresos de explotación/Gastos de explotación	4
Pasivo líquido/Activo Total	3
Existencias/Ingresos de explotación	3
Fondos propios/ Activo total	3
Deuda total/Fondos propios	3

Tabla 2.2: Variables usadas en diversos estudios desde 1966 hasta 2009

2.1.3. Modelos de resolución

En este punto vamos a hacer una breve introducción histórica sobre los modelos usados para la resolución y posteriormente hacer una mención a cada uno de los modelos más utilizados. No haremos demasiado énfasis en la explicación de estos modelos, puesto que los modelos más interesantes se explicarán en profundidad en el siguiente punto.

Contexto histórico.

En el caso de las técnicas usadas para la resolución de este problema podemos observar como en la tabla 6 de ([Tascón and Castaño, 2010](#)), etiqueta diferentes estudios a lo largo de los años, según el modelo matemático usado para la resolución. Es fácil observar que lo utilizado en las primeras investigaciones son análisis univariante y análisis discriminante múltiple que representan modelos lineales muy sencillos para problemas de clasificación.

Pero es a partir a mediados de los años 70 cuando comenzamos a ver el uso de modelos como la regresión logística, que es un modelo de discriminación lineal como los anteriormente mencionados, y que no presenta ciertas restricciones a la hora de aplicarlos en ciertas variables como explicaremos posteriormente. Además, también por esta década aparecen el uso modelos de análisis de probabilidad condicional para este tipo de estudios.

En un principio no hay diferencias significativas entre los resultados aportados por los modelos de análisis univariante y discriminante múltiple, y los modelos regresión logística y probabilidad condicional, pero si hay estudios como el de ([Lennox, 1999](#)), que hace una reevaluación de los modelos de regresión logística y de probabilidad condicional, donde si se obtienen unos mejores resultados cuando la especificación de las variables mejora.

Es con la revolución de la computación y de la inteligencia artificial, a partir de los años 90, donde se introducen técnicas de inteligencia artificial para la resolución de este tipo de problemas, y se comienzan a usar modelos como las redes neuronales que es un modelo que se desarrolla por primera vez con ([Rosenblatt, 1958](#)) o los arboles de decisión. Añadir que este tipo de modelos resultan modelos bastante potentes a la hora de aportar nuevas soluciones a estos problemas porque son capaces de resolver problemas de clasificación no lineales lo cual hace se sienta un soplo de aire fresco en este tipo de investigaciones.

Análisis Univariante.

El análisis univariante lo introduce (Beaver, 1966) para la resolución de problemas como el fracaso empresarial, y se toma como trabajo de referencia. Este tipo de análisis, suele estar presente en el comienzo de cualquier investigación, porque mediante medidas estadísticas de tendencias centrales y estadísticos de dispersión pueden ver como se distribuye esa variable a lo largo del conjunto de datos. Posteriormente, mediante técnicas de inferencia estadística obtener la significatividad que tiene cada una de las variables para predecir la clasificación a una empresa como fracasada o no.

El procedimiento que se suele seguir para este tipo de técnica está representado en la figura 2.1

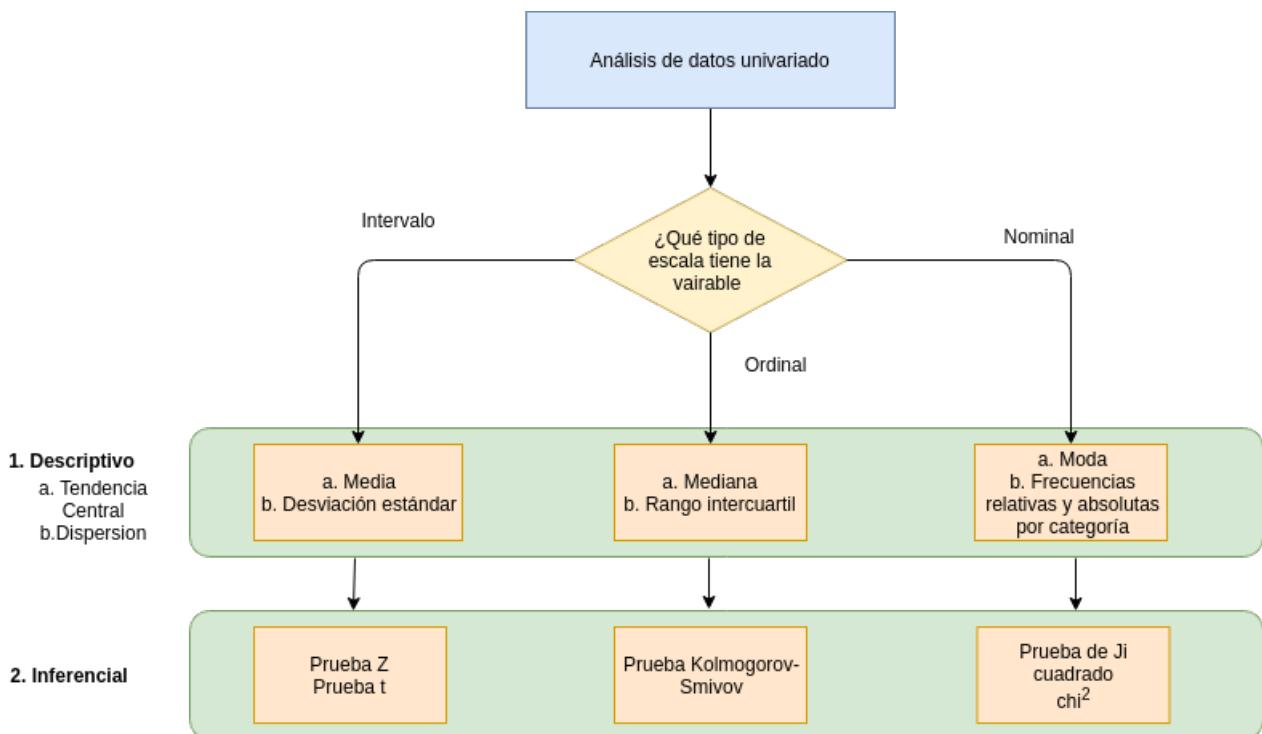


Figura 2.1: Procedimiento del análisis univariado.

Análisis Multivariante Discriminante.

En cuanto al análisis multivariante discriminante es una técnica estadística que tiene como objetivo analizar si existen diferencias significativas entre la variable clase y un conjunto de variables medidas sobre esa variable clase. En caso de existir esas diferencias, pueden facilitar modelos de clasificación automática de nuevas instancias. Con este tipo de técnicas multivariantes se pretenden encontrar correlaciones lineales entre las variables continuas que mejor discriminan a la variable clase dada ([Fernández, 2011](#))³.

La ventaja que obtiene el análisis discriminante multivariante sobre los modelos univariados, es que hay variables que descartaríamos en el análisis univariado porque no aportan información significativa al problema, que en un análisis multivariante discriminante esas mismas variables descartadas junto con otras variables si puedan aportar información significativa al problema.

Regresión Logística.

La regresión logística es un tipo de regresión lineal que se utiliza para predecir una variable categórica, en función de un conjunto de variables predictoras. Es por tanto, un tipo de modelo matemático que en caso de tener como variable clase, una variable dicotómica, estimar la probabilidad de que se de ese uno de esos sucesos. ([Jimmy Reyes Rocabado, 2007](#))⁴.

Aunque si nos damos cuenta no tendría mucho sentido puesto que una regresión lineal como tal nos daría una función como en la figura 2.2, para hacer la clasificación.

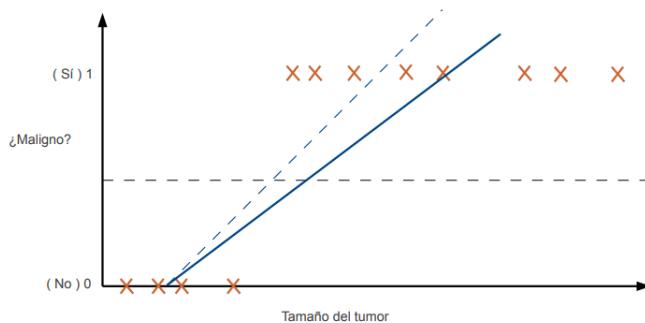


Figura 2.2: Regresión Lineal para clasificación

³<http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/DISCIMINANTE/analisis-discriminante.pdf>

⁴<https://scielo.conicyt.cl/pdf/estped/v33n2/art06.pdf>

A medida que se aumenten el numero ejemplos la hipótesis de la regresión lineal, puede dejar de ser valida por eso se necesita otro tipo de modelo matemático que podamos usar a partir de una regresión lineal que pueda modelar la clasificación.

En regresión logística, $y \in \{0,1\}$, por lo que se busca que $0 \leq h_\theta(x) \leq 1$, donde $h_\theta(x)$ es el resultado de aplicar la función logística o sigmoidal $\theta^T x$:

$$h_\theta = \frac{1}{1 + e^{-\theta^T x}}$$

Ahora la función resultante figura 2.3 se adapta mejor al problema a resolver:

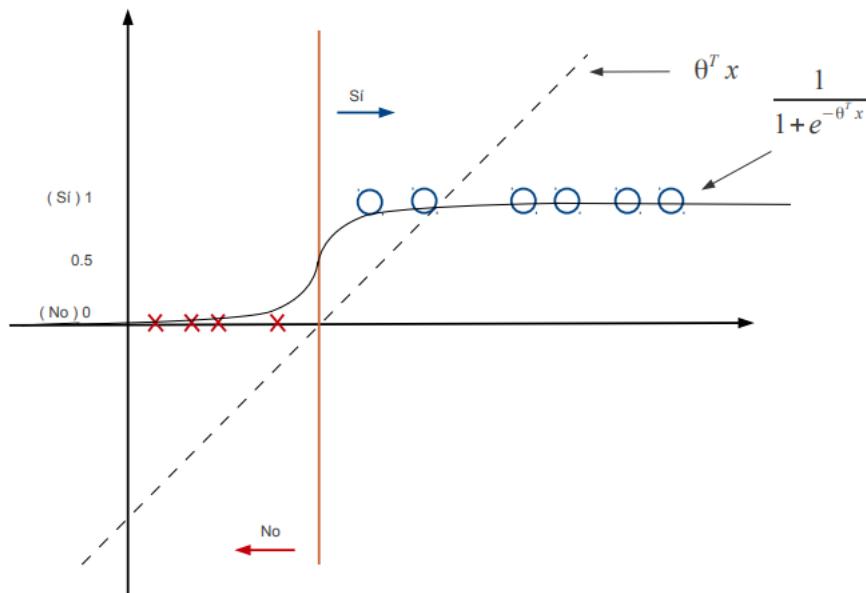


Figura 2.3: Regresión Logística para clasificación

Decir que aunque la regresión logística y el análisis discriminante son clasificadores lineales, la regresión logística posee ventajas al ser mas flexible que el análisis discriminante, como por ejemplo, que no requiere de variables que se distribuyan de una manera normal.

Modelos de probabilidad condicionada.

Los modelos de probabilidad condicionada son aquellos que su modelo matemático esta basado en la probabilidad condicionada calculada usando el teorema de Bayes. De hecho, el modelo más sencillo es Naive Bayes, donde la hipótesis se basa en la independencia probabilística. Donde, variables predictoras se consideran condicionalmente independientes conocido el valor de la variable clase:

$$I(X_i|Y|S), \forall S \subseteq \{X_1, X_2, \dots, X_i, X_{i+1}, \dots, X_n\}$$

Gráficamente se vería como en la figura 2.4.

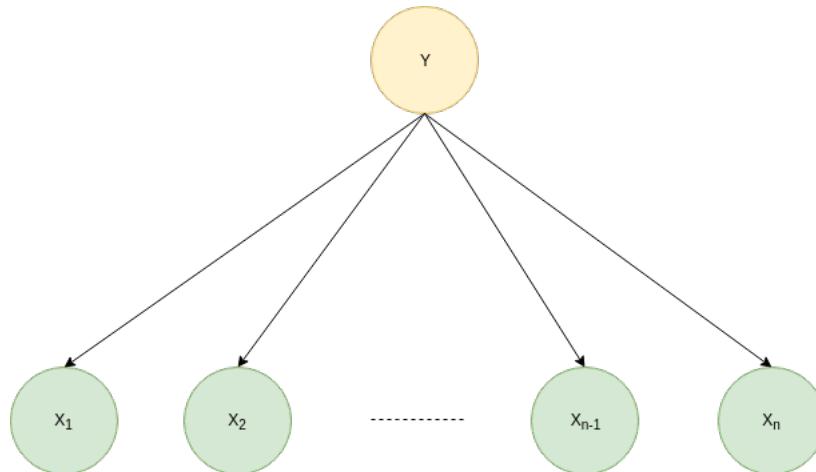


Figura 2.4: Modelo Naive Bayes

Este, aunque es un modelo muy sencillo, el tiempo computacional de Naive Bayes para el entrenamiento es de $O(n)$, es decir, muy barato y puede llegar a dar muy buenos resultados. Existen, modelos más complejos basados en las redes Bayesianas sin restricciones, pero que tienen una complejidad exponencial en el aprendizaje, lo cual lo hace inviable para cantidades de datos grandes. Por ello, clasificadores como el TAN(Tree Augmented Naive Bayes)⁵, que es un modelo a medio camino entre el modelo Naive Bayes y las redes Bayesianas suelen ser la mejor opción.

Árboles de decisión.

Un árbol de decisión o, como lo denominan en el trabajo de (Tascón and Castaño, 2010), algoritmo de particiones recursivas. Es una función de hipótesis mediante un grafo dirigido que cumplen estas normas:

- Cada nodo representa una variable predictora.
- Cada rama que pende de un nodo representa uno de los posibles valores que puede tomar la variable correspondiente.
- Las hojas corresponden con valores de la variable clase.
- Un ejemplo se clasifica recorriendo el árbol dese la raíz y eligiendo cada momento la rama que satisface la condición para el valor del atributo correspondiente. La clase asignada a la hoja a la que se llega.

⁵<https://www.sciencedirect.com/science/article/abs/pii/S0950705109000033>

Gráficamente tienen el aspecto de la figura 2.5:

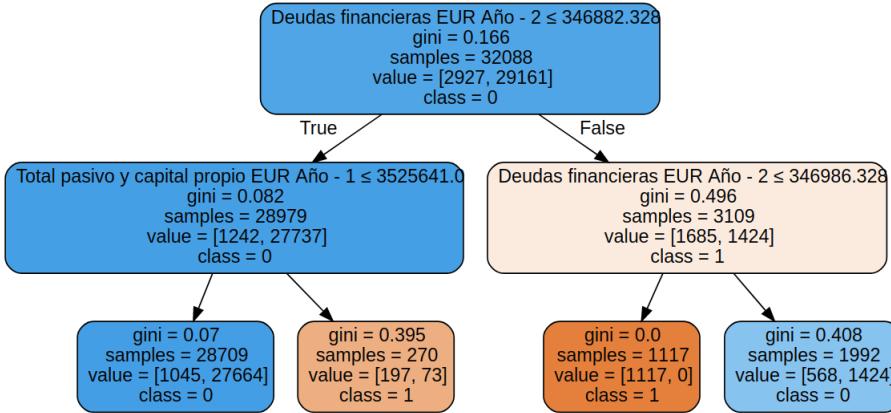


Figura 2.5: Árbol de decisión

Comentar que los modelos, hasta este momento son los modelos más sencillos y más fáciles de entender, pues son la base de los modelos explicados en el siguiente capítulo, que son los más potentes y los que hasta la fecha de hoy están dando mejores resultados.

2.2. Hot topics en Aprendizaje Automático.

En esta sección vamos a hablar sobre las técnicas o modelos del aprendizaje automático, más potentes que se están utilizando para dar solución al tipo de problema que nos estamos enfrentando en este Trabajo Final de Grado y a muchos otros. Para ello vamos a hacer una breve introducción sobre el campo del Aprendizaje automático, para poner en contexto el dominio del problema sobre el que nos estamos moviendo.

2.2.1. Introducción. ¿Qué es el aprendizaje automático?. ¿Hay alguna diferencia con la inteligencia artificial?.

En estos últimos años, parece que la palabra inteligencia artificial y aprendizaje automático o "Machine Learning" se ha convertido en el mismo término y la realidad es que esto no es así. Para poder definirlo y contextualizarlo de la manera correcta, lo primero que tendremos que hacer es definir la palabra inteligencia. Anticipar que este no es una palabra que contiene una definición única:

- Según (Mainstream, 1994)⁶: 'Una capacidad mental muy general que, entre otras

⁶<http://www.intelligence.martinsewell.com/Gottfredson1997.pdf>

cosas, implica la habilidad de razonar, planear, resolver problemas, pensar de manera abstracta, comprender ideas complejas, aprender rápidamente y aprender de la experiencia. No es un mero aprendizaje de los libros, ni una habilidad estrictamente académica, ni un talento para superar pruebas. Más bien, el concepto se refiere a la capacidad de comprender el propio entorno.”

- Universidad de Cuenca ([Viviana Janeth, 2010](#)): “la capacidad cerebral por la cual conseguimos penetrar en la comprensión de las cosas eligiendo el mejor camino. ”
- ([Google, 2019](#))⁷ :“Habilidad o capacidad para hacer algo con facilidad, acierto y rapidez.”
- ([RAE, 2019](#))⁸ :“Capacidad de entender o comprender. Habilidad, destreza y experiencia.”

Como podemos observar todas estas definiciones, formulan un concepto de inteligencia diferente entre ellos. Vamos a intentar extraer una idea común. Entonces, la inteligencia es la habilidad de la mente humana que permite resolver problemas, tomar decisiones, y aprender nuevos conceptos a partir de experiencias previas o información que es capaz de transformar en conocimiento.

Según acuña John McCarthy, en 1956 por primera vez en la Conferencia de Dartmouth, la inteligencia artificial es una disciplina del campo de la Informática, que busca la creación de máquinas que puedan **imitar** comportamientos inteligentes. Lo que quiere decir que comportamientos como el que tiene un brazo robótico para montar las diferentes piezas de un coche se puede programar de manera clásica y la maquina no tiene la necesidad de “aprender”, y en cambio podríamos decir que se trata de un comportamiento inteligente si tomamos la definición de Google como definición de inteligencia, pero que no entraría en el concepto de aprendizaje automático como definimos en el capítulo de introducción.

También mencionar que existen dos tipos de inteligencia artificial ([Hardy, 2001](#))⁹:

- Inteligencia artificial débil: Es aquella que es capaz de **simular** estados mentales(sin ser estados mentales), de nuestro cerebro por medio de computadores, de una tarea específica.

⁷<https://www.google.com/search?q=concepto+de+inteligencia&oq=concepto+de+inteligencia&aqs=chrome..69i57j0j69i61l2j0l2.3144j1j4&sourceid=chrome&ie=UTF-8>

⁸<https://dle.rae.es/?id=LqtyoaQ|LqusWqH>

⁹<https://www.redalyc.org/pdf/305/30500219.pdf>

- Inteligencia artificial fuerte: Este campo de la inteligencia artificial sostiene que las máquinas son capaces de realizar maquinas realmente pensantes y con estados mentales propios.

De esta manera los conceptos introducidos como conjuntos tendrían la forma de la figura 2.6. Donde, nos damos cuenta que el estado del arte se encuentra en la inteligencia artificial débil, en concreto en el aprendizaje automático, que dista mucho de crear estados mentales propios por parte de las computadoras.

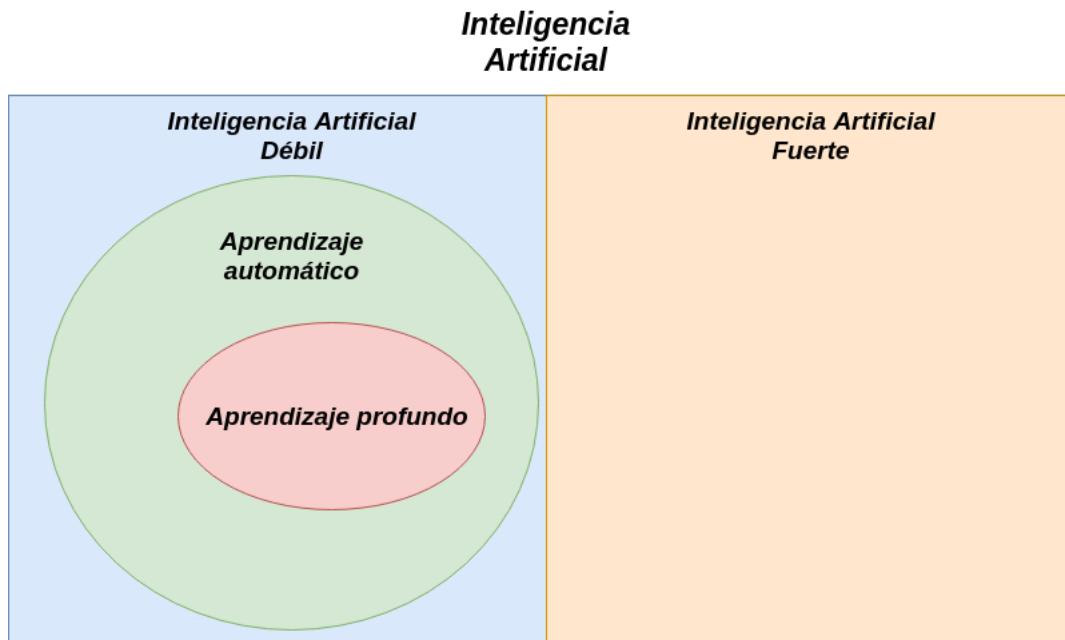


Figura 2.6: Inteligencia artificial conjunto

2.2.2. Redes Neuronales y aprendizaje profundo.

Es con la llegada de la digitalización en los años 90, y con la recopilación masiva de datos, donde el aprendizaje automático comienza otra vez a resurgir como área del conocimiento. Pero no ha sido hasta la última década cuando la cantidad de datos se ha ido duplicando año a año según publican portales de noticias online como ([Tilves, 2017](#))¹⁰, donde apunta que la cantidad de datos existentes en 2017, se habría creado desde el periodo de 2015 a 2017. Es ahí, donde aparecen las redes neuronales y el aprendizaje profundo.

Una red neuronal no es más que una abstracción matemática de como funcionaría una neurona biológica como apunta ([Rosenblatt, 1958](#)) por primera vez y que tiene la

¹⁰<https://www.silicon.es/datos-infografia-2333354>

forma de la figura 2.7. Un perceptrón es el modelo matemático más simple de lo que podría ser una neurona biológica, donde como podemos ver en la figura 2.7, tenemos una neurona que es la encargada de tener la unión sumadora para posteriormente aplicarle una función de activación en este caso puede ser la sigmoide. Las dendritas serían las aristas del grafo dirigido donde se encuentran los pesos sinápticos donde se produciría la sinapsis con la entradas, y el canal del salida que es el axón. Este modelo es muy simple y solo es capaz de separar con un hiperplano los elementos a clasificar. Este ejemplo de perceptrón no produce, otra cosa que una regresión logística. Si nos damos cuenta la unión sumadora da como resultado:

$$\theta_1x_1 + \theta_2x_2 + \theta_3x_3 + \theta_4x_4 + \theta_5x_5 = \theta^T x$$

Y si aplicamos la función de activación a la salida de la neurona da como resultado la función sigmoide de nuevo:

$$h_\theta x = \frac{1}{1 + e^{-\theta^T x}}$$

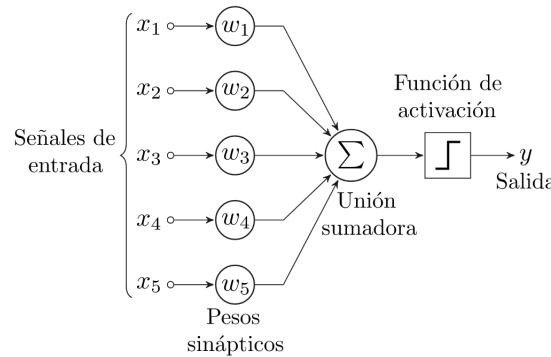


Figura 2.7: Perceptron propuesto por Rosenblatt

Hemos de mencionar que esta es la forma más primitiva de las redes neuronales, ahora bien, las redes neuronales actuales están formadas por varias capas de neuronas artificiales que tienen:

- 1 capa de entrada
- 0 (el caso del perceptrón) o más capas ocultas.
- 1 capa de salida.

De esta manera las redes neuronales son capaces de aprender manera jerarquizada características del problema para luego resolverlo en su conjunto, es decir, que es capaz de aprender una representación de los datos con diferentes niveles de abstracción. Esto se ve perfectamente en la figura 2.8, donde para resolverlo tienes que visualizar que

forma gráfica tenía el problema de clasificación y a partir de ahí, darte cuenta que la manera de separarlo es mediante una circunferencia, la cual se puede describir como $x_1^2 + x_2^2$. Pero para resolverlo se tiene que hacer un procesamiento a mano extracción de características, exportarlas como nuevo input, y finalmente aplicarle un algoritmo de aprendizaje automático como una regresión logística. Lo que ocurre es que este tipo de procedimiento cuando la cantidad de variables no son dos si no miles es un proceso totalmente impracticable manualmente.

Es en este momento donde entran las redes neuronales y lo que se conoce como aprendizaje profundo, un ejemplo de juguete como el anterior se puede resolver con una red neuronal simple, sin necesidad de hacer un proceso de extracción de características tradicional. Simplemente una configuración como la de la figura 2.9, es capaz de resolver un problema de clasificación que a priori no se podía resolver sin una extracción de características.

Lo que ha hecho la red neuronal ha sido hacer una separación con diferentes cortes de manera lineal del problema y “fusionar” esas linealidades para formar una nueva característica no lineal, de tal manera, que el resultado fuera prácticamente una circunferencia y clasificar perfectamente ese problema. De hecho, la ultima capa oculta nos la podríamos haber ahorrado pues la neurona de salida se hubiera encargado de separarlo también, pero hemos preferido aumentar una capa de la red para que suavizara aún más a la circunferencia.

Este ejemplo de juguete se puede extrapolar a ejemplos mucho mas complejos como la clasificación de imágenes, genómica, procesamiento del lenguaje natural, etc.

Aunque, no lo hemos mencionado anteriormente, para que estos modelos funcionen la clave es que los pesos de la red neuronal estén bien ajustados, esto lo hace mediante los mecanismos de propagación hacia adelante y propagación hacia atrás:

- Propagación hacia adelante: La propagación hacia adelante consiste en ir propagando las entradas a través de la red neuronal como en el ejemplo del perceptrón expuesto al principio de este punto. La única diferencia es que la salida de la función de activación de esa neurona será la entrada de la siguiente. Así hasta llegar a la última capa.
- Propagación hacia atrás: Una vez se ha propagado un ejemplo y se conoce su valor predicho, este se compara con el valor real (conocido para las instancias de entrenamiento) y se calcula el error cometido. El algoritmo de retro-propagación

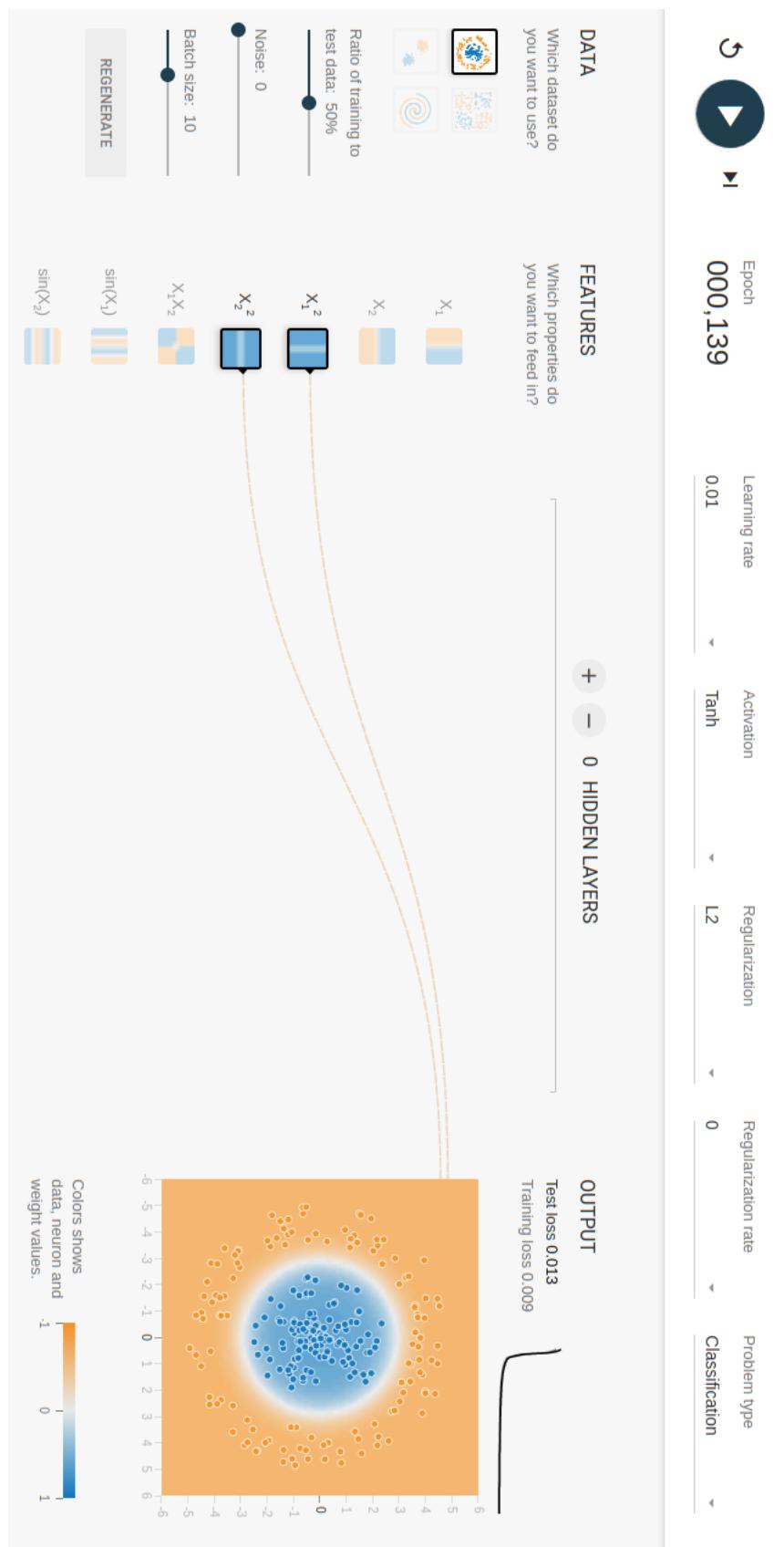


Figura 2.8: Problema de clasificación resuelto de manera tradicional.

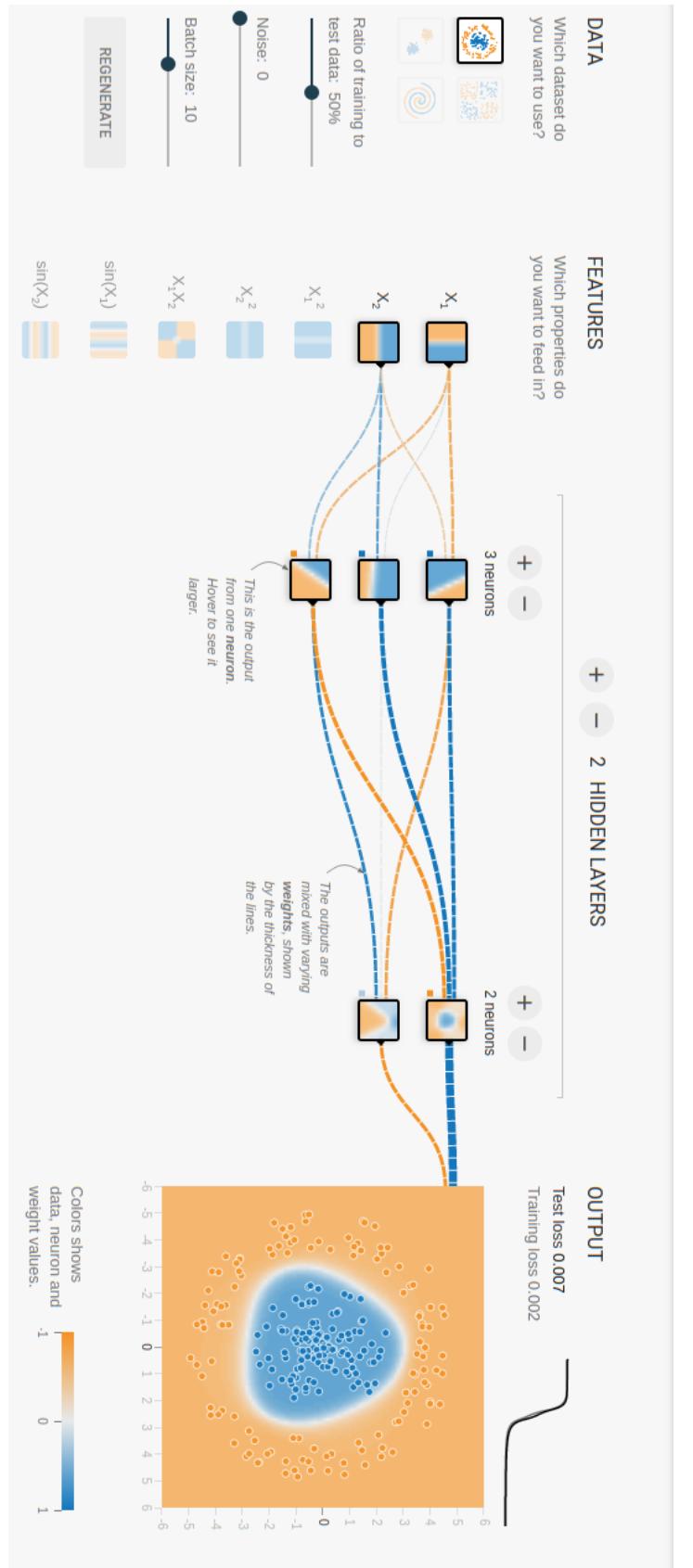


Figura 2.9: Problema de clasificación resuelto con un red neuronal.

usa el algoritmo de descenso del gradiente (Mamood, 2019)¹¹ para optimizar la función de de coste o minimización del error, propagando el error hacia los niveles anteriores de la red y re-ajustando el vector de pesos para minimizar el error cometido. Normalmente se necesitan muchas iteraciones para que los pesos converjan a valores cercanos al óptimo.

En nuestro caso de estudio como es el del fracaso empresarial, las redes neuronales han sido de uso reiterativo desde comienzos de los noventa. Autores como Fletcher, Goss, Wilson y Sharda, Altman han utilizado este modelo matemático para análisis discriminativos de este problema, en muchos de ellos los resultados eran significativamente mejores que con modelos lineales. Aunque hay que apuntar que en muchos de ellos como es en el caso de (Altman, 1994), donde hace una comparación entre las redes neuronales y los discriminantes lineales como el multivariable y la regresión logística sin encontrar, diferencias significativas. Esto tiene una explicación muy sencilla, y es que el algoritmo de propagación hacia atrás necesita un número alto de instancias para poder identificar esas relaciones no lineales entre las variables, además que en muchas ocasiones acertar el número de capas y neuronas por capa puede resultar un trabajo bastante empírico puesto que no hay una teoría firme sobre este problema relacionado con las redes neuronales.

Mencionar, que además este tipo de modelos son de caja negra, es decir, que los pesos asociados a su arquitectura una vez la red esta entrenada, no son posibles de explicar, por ello en muchos estudios de este tipo los investigadores son reacios a usar este tipo de técnicas, puesto que el objetivo hoy por hoy es obtener algún modelo matemático que sea capaz de dar una teoría robusta de porque se produce el fracaso empresarial en una empresa.

2.2.3. Redes neuronales recurrentes. LSTM(Long Short Term Memory).

La idea inicial de las RNNs es simular de alguna manera las funciones como las harían un cerebro humano. Pero si nos damos cuenta el cerebro humano no funciona olvidando todo desde cero y lo piensa otra vez. En cambio, esto es lo que hacen las RNNs clásicas y es una deficiencia muy importante.

Un ejemplo, si a nosotros nos pusieran el problema de seguir con la siguiente palabra después de una frase, sería muy difícil de adivinar sin el contexto de la frase. Justo este problema es resuelto con este tipo de redes neuronales recurrentes. Este tipo de redes

¹¹<https://towardsdatascience.com/gradient-descent-3a7db7520711>

son aquellas donde la salida pasa como entrada de la siguiente iteración. La intuición gráfica sería como la de la figura 2.10. Con este simple cambio lo que consigue la red neuronal es que el resultado anterior pase a formar parte de las variables de entrada de la nueva iteración, contextualizando el problema.

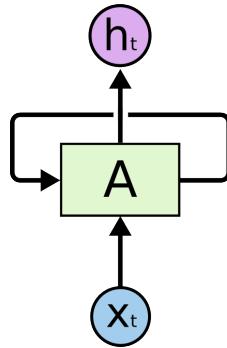


Figura 2.10: Red neuronal recurrente básica.

Aunque parezca que el paradigma de la RNNs clásicas ha cambiado de manera drástica, si nosotros desenrollamos el bucle que le hemos añadido a esta RNN como se ve en la figura 2.11, lo único que se le está haciendo es de alguna manera hacer muchas copias de la estructura de una misma RNN y e ir pasando la salida de cada una a su sucesora.

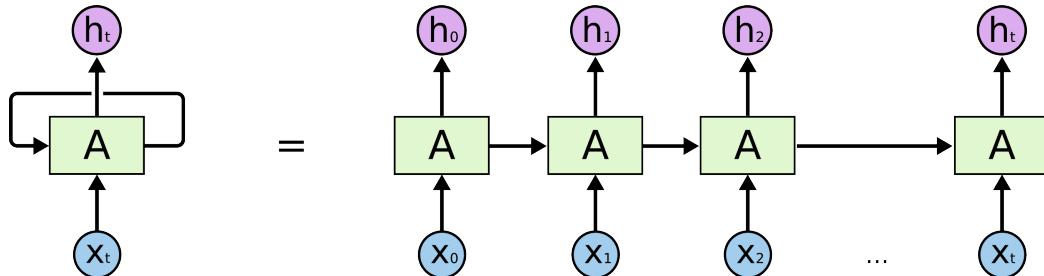


Figura 2.11: Red neuronal recurrente básica desenrollada.

Parece que con esta arquitectura ya está el problema que planteábamos al principio, resuelto. Ahora imaginemos que, el contexto que se quiere recordar no estaba ni en la misma frase para predecir la siguiente la palabra correctamente si no que se encuentra mucho más separado como por ejemplo, “el dia de ayer hacía un sol esplendoroso... tras salir de mi casa me di cuenta que hacía mucho *calor*”. Para casos como el anterior la RNN no tiene el rendimiento esperado porque no tiene la capacidad conectar información tan separada en la serie temporal. De ahí nacen las LSTMs.

Redes LSTM

Las Long Short Term Memory, son aquellas RNN que son capaces de aprender a contextualizar o aprender las dependencias en series temporales muy largas y son introducidas por (Hochreiter and Schmidhuber, 1997)¹², popularizándose de manera muy rápida y aportando grandes contribuciones, sobre todo en el campo del Procesamiento del Lenguaje.

La estructura de la red neuronal recurrente que hemos utilizado como primera aproximación sería como en la figura 2.12, donde simplemente concatenamos la salida de la antecesora como entrada de la sucesora. Una arquitectura muy sencilla.

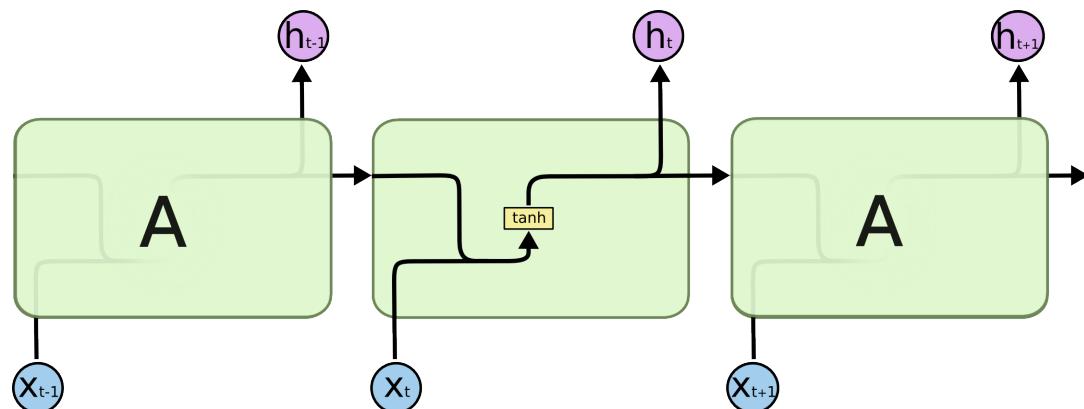


Figura 2.12: Red neuronal recurrente interna básica desenrollada.

Sin embargo, en el caso de la neurona A de una LSTM, tiene una arquitectura como la de la figura 2.13, donde se puede observar que la arquitectura es un poco más especial.

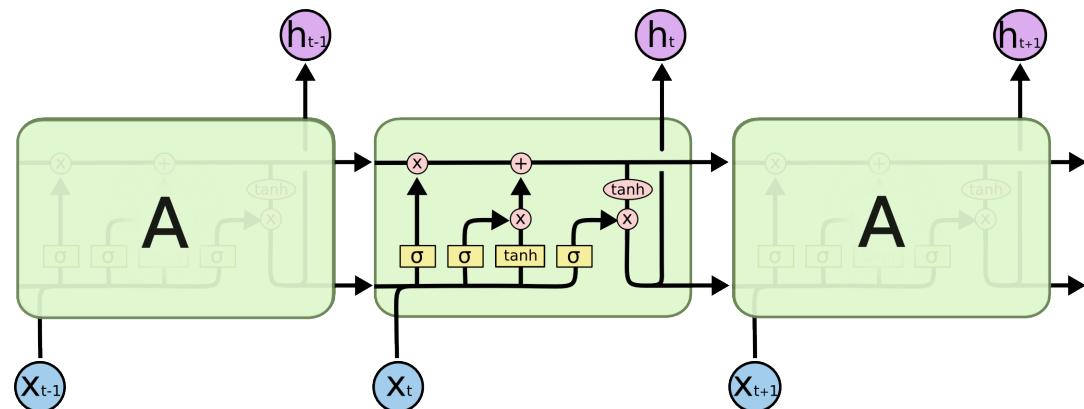


Figura 2.13: LSTM desenrollada.

¹²<http://www.bioinf.jku.at/publications/older/2604.pdf>

De manera muy resumida, lo que se consigue con esta arquitectura de neurona es que por el flujo de información que pasa desde C_{t-1} a C_t la información de la entrada anterior fluya por la neurona prácticamente sin cambios, y las funciones de activación sigmoideas sean las que se encarguen de producir números entre 0 y 1, que describirán la cantidad que debe pasar de cada componente. Para ahondar más en las matemáticas que hay en las LSTM recomiendo este enlace ([Olah, 2015](#))¹³.

En concreto hemos explicado este modelo de RNN, porque se ha implementado en problemas de procesamiento del lenguaje, predicción del stock, etc. Y se espera que pueda tener una buena aplicación en nuestro problema que será explicado como y porque se ha implementado en este problema en concreto.

2.2.4. Combinacion de clasificadores (Ensembles)

Un ensemble es un conjunto de clasificadores que se entrena a partir de los mismos datos de entrenamiento, para a posteriormente que cada uno clasifique de manera individual y estos luego agreguen sus predicciones para llegar a la predicción final.

Este tipo de modelo suele mejorar respecto del clasificador individual, aunque con el inconveniente como pasaba en el caso de las RNN, son difíciles o casi imposible de analizar.

Bagging

Bagging, también conocido como Bootstrap Aggregating, es una combinación de clasificadores como puede ser los árboles de decisión que tiene como objetivo reducir la varianza (que es debido al sobreajuste) de los modelos al computar la moda o hacer medias en predicción numérica. Aunque habitualmente un mayor número de clasificadores implica un mejor rendimiento, también aumentan las necesidades de tiempo y espacio. El bagging consta de las siguientes fases:

- Fase de entrenamiento: Dado un subconjunto D de n ejemplos, en cada iteración t se obtendrá un nuevo conjunto D_t , del mismo tamaño que el anterior, mediante un muestreo aleatorio con reemplazo de D . Posteriormente se aprenderá un modelo de clasificación M_t a partir de cada conjunto D_t .
- Fase de inferencia: Cada uno de los clasificadores clasifica el nuevo ejemplo entrante, una vez que todos los modelos han clasificado el nuevo ejemplo devuelve

¹³<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

la clase más frecuente, es decir, la moda. En caso de ser un problema de regresión se haría el análogo, es decir, la media.

Este tipo de modelos funcionan muy bien con algoritmos que tienen un error debido a la varianza, como puede ser un árbol de decisión con una gran profundidad, y que pequeños cambios en los datos de entrada pueden producir cambios importantes en la clasificación. El efecto gráfico que tendría sobre el resultado sería como el de la figura 2.14, donde en la primera imagen podríamos ver el problema real, es decir, como se distribuyen los datos del problema. La segunda imagen se vería como un árbol de decisión podría clasificar ese problema. Y por último como al usar un bagging de árboles se es capaz de trazar una frontera más difusa entre las clases que no produzca ese sobreajuste.

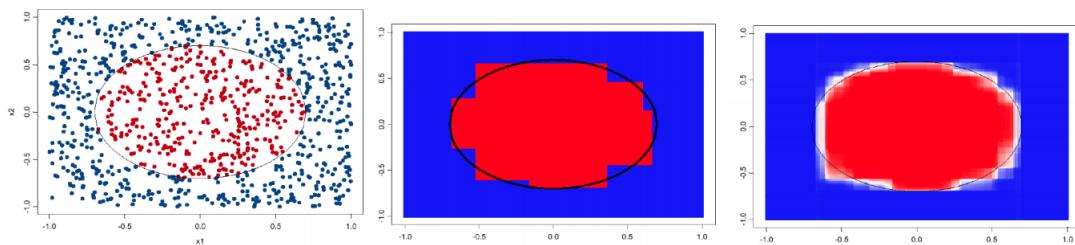


Figura 2.14: Efecto del bagging en la clasificación

Bosques aleatorios. (Random Forest)

El modelo de Bosque Aleatorio también conocido como Random Forest, es una técnica de bagging donde la combinación de modelos es en forma de árbol.

El proceso de construcción de estos arboles es de manera semi-aleatoria, lo que quiere decir es que en cada paso de la construcción del árbol, de un conjunto de variables p se toma un subconjunto de variables de manera aleatoria m , donde el número de variables en m es mucho menor que en p . Posteriormente se evalúan, y se escogen las mejores variables de ese subconjunto m para formar un nodo. Este procedimiento se hace para cada árbol generado.

Tiene ciertas ventajas frente a otros modelos de combinación de modelos. Puesto que, por la semi-aleatorización de la selección de las variables, ahorra bastante tiempo en el proceso de construcción de los árboles, ya que debe de evaluar todo el conjunto de datos, si no un subconjunto de variables bastante menor. Además, no se produce overfitting cuando el número de árboles es alto, y obtienen en precisión resultados muy

parecidos al boosting aunque no cambien el conjunto de entrenamiento de manera progresiva.

Boosting

El boosting también es una técnica de ensembles, que tiene como objetivo reducir el error debido al sesgo, el boosting tiende a lograr una mayor precisión que el bagging, pero a veces sobreajusta a los datos mal clasificados, lo que le puede hacer sensible al ruido. El boosting consta de las siguientes fases:

- Fase de entrenamiento: Se asignan pesos a cada ejemplo del conjunto entrenamiento D , y se aprende de manera iterativa, un conjunto de clasificadores T . Donde para el aprendizaje del modelo M_t se usan los pesos asociados a las instancias, para posteriormente aquellas instancias mal clasificadas se les incrementarán esos pesos y el modelo M_{t+1} aprenderá con los nuevos pesos. De esta manera en cada iteración se aprende un modelo que presta más atención a los casos anteriormente mal clasificados.
- Fase de inferencia: Cuando hay una nueva instancia el modelo final M^* , donde cada clasificador individual clasifica esa instancia haciendo una ponderación de cada clasificador M_t en función de su precisión.

Gráficamente el algoritmo de boosting haría como en la figura 2.15. Donde, de manera iterativa se han creado 3 árboles de profundidad 1 que separan el espacio de clasificación con una precisión bastante mala, pero cuando se crea el modelo M^* es capaz de en entrenamiento tener una precisión del 100 %. Comentar que el modelo base era un modelo tan básico que su error era debido al sesgo pero con el algoritmo de boosting ha sido capaz de solventar perfectamente ese problema, esto si se sobreentrena puede llegar a sobreajustarse y que el error sea debido a la varianza.

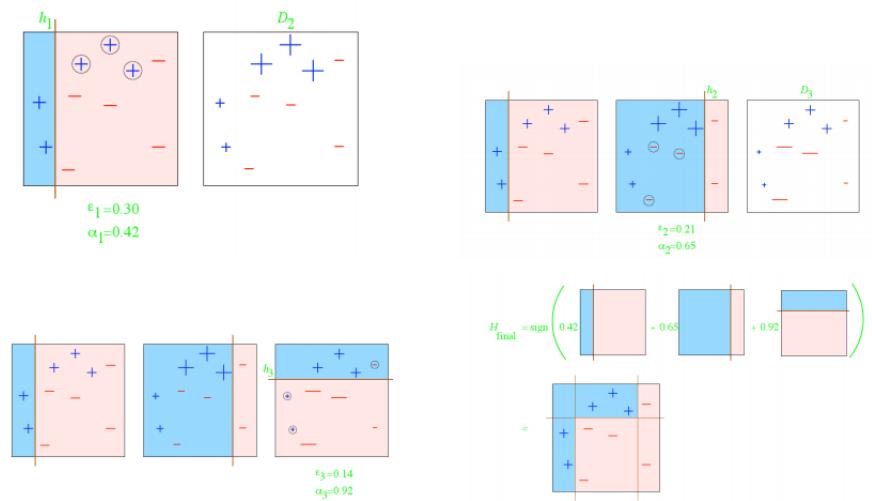


Figura 2.15: Efecto del boosting en la clasificación

Capítulo 3

METODOLOGÍA Y DESARROLLO

En este capítulo, describiremos cual será la metodología que se usa en este proyecto de minería de datos, y de que fases consta.

3.1. Descripción de la metodología

La metodología elegida para llevar a cabo este TFG será CRISP-DM(Cross-Industry Standard Process for Data Mining) con la finalidad de poder seguir un proceso estandarizado que no este asociado a ningún tipo de tecnología ni herramienta. Esta nos proporcionara un marco de trabajo para guiar el proceso de minería de datos. Esta metodología lleva a cabo las diferentes fases:

- Comprensión del negocio: Esta fase del proyecto se basa en la definición de las necesidades del cliente y la obtención de conocimiento del dominio del problema. Su finalidad es marcar unos objetivos claros antes para poder ser capaces de optimizar el tiempo que lleva a cabo cada fase del proyecto de la manera más organizada posible. Como en este caso no contamos con un cliente con el que lidiar para satisfacer esas necesidades, sera el alumno y el director el TFG los que elegiremos cuales son los objetivos principales del proyecto.
- Estudio y compresión de los datos: Esta parte del proyecto comienza con la obtención en bruto de todos los datos iniciales y su posterior estudio tanto para familiarizarnos con los datos, como para identificar si los datos con los que estamos tratando están incompletos o tienen problemas de ruido e inconsistencia. Además esta fase nos puede ayudar a detectar subconjuntos de datos que apoyen nuestra hipótesis inicial.
- Análisis de los datos, selección y transformación de las variables predictoras: Esta puede que se trate de una de las partes más sensibles del proyecto puesto que el proceso llevado en la selección, limpieza y construcción de los datos en bruto puede determinar de manera muy drástica lo buenos o malos que van a ser los modelos implementados en fases posteriores.
- Modelado: Llegados a este punto se eligen las técnicas de modelado y su parametrización óptima. En muchas ocasiones estas técnicas poseen características propias de los datos que son necesarias para su implementación por ello no es de extrañar que se deba de volver a la fase anterior.
- Evaluación: Una vez construidos diferentes modelos para el conjunto de datos final, procedemos a evaluarlos. Esta fase es de gran importancia pues de aquí saldrá el modelo que mejor generalice y podremos sacar diferentes conclusiones de si alguna decisión tomada en etapas anteriores ha de ser corregida.
- Puesta en producción: Cuando elegimos uno de los modelos obtenidos el proceso de minería de datos puede ser que no acabe aquí y se deba hacer otra iteración de las diferentes etapas seguidas hasta llegar a este punto otra vez.

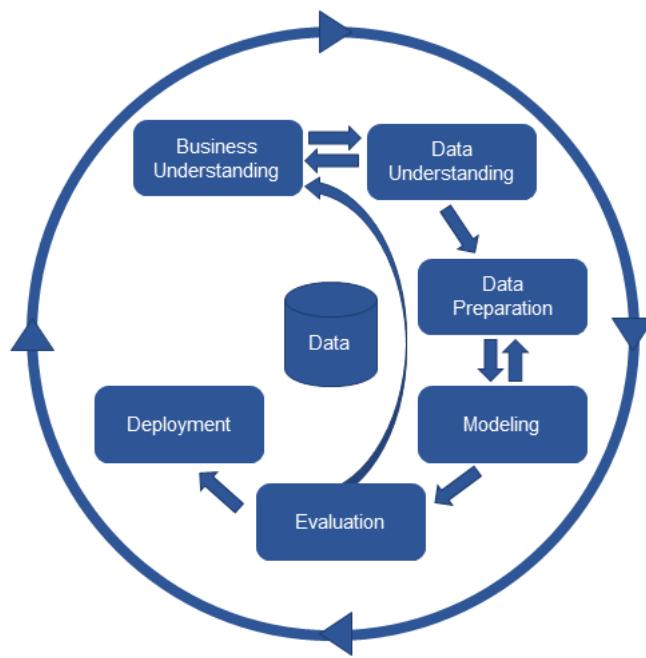


Figura 3.1: Proceso completo de minería de datos.

Si nos fijamos en la Figura 3.1. El proceso de minería de datos no se hace de manera lineal si no que podemos volver a etapas anteriores si esto fuera necesario. En cuanto al tiempo estipulado para cada etapa podemos decir que las etapas hasta llegar al modelado podrían ocupar un 70/80 % de la horas totales del TFG y el otro 20 % restante corresponde a las ultimas dos etapas del proceso CRISP-DM.

Capítulo 4

Fases de trabajo

Dado que la metodología CRISP-DM puede requerir varias iteraciones, se hará una nueva sección para cada una de las iteraciones realizadas en este proyecto. En cada una de las secciones describiremos el trabajo realizado para cada una de las iteraciones.

4.1. Primera iteración

4.1.1. Comprensión del negocio

El fracaso empresarial es un campo del conocimiento, donde de manera constante se intenta mejorar los modelos predictivos y donde se invierte mucho dinero con el fin de tomar los menores riesgos posibles en el mundo de los negocios. Aquí es donde entra en juego la minería de datos y el aprendizaje automático, estas disciplinas han demostrado ser una de las mejores soluciones a este problema.

No obstante, antes de comenzar con el proceso de minería de datos sería mas importante comenzar a obtener comprender el dominio del problema del mundo empresarial y las ciencias sociales. Comenzaremos con la definición de fracaso empresarial y que tipos existen de este.

No existe una teoría firme sobre el mismo, sino que por el contrario, se trata de un concepto amplio que engloba una diversidad de estados que pueden repercutir de forma negativa en la empresa, siendo también varias las causas por las que una empresa entra en crisis, así como los síntomas de deterioro que se pueden apreciar en ella ([Gallego, 2004](#)).

Ahora si, el fracaso empresarial se puede distinguir en 3 tipos:

- Fracaso económico: Según ([Altman, 1968](#)) el fracaso económico de una empresa surge cuando estas comienzan a recibir una rentabilidad económica por sus inversiones, menor que las del mercado, ante una situación de riesgo. Es decir, cuando los activos de la empresa son insuficientes para cubrir los pasivos, o lo que es lo mismo la empresa no es capaz de cubrir su deuda con los ingresos obtenidos. Esto es algo que puede ocurrir sobre todo en empresas jóvenes. El problema ocurre cuando la perdida de liquidez se mantiene durante el tiempo y la empresa no es capaz de hacer frente a los pagos de sus acreedores. Esto es provocado por unas constantes pérdidas o por una débil estructura financiera, lo que lleva a la empresa a la suspensión de pagos de los trabajadores y posteriormente a la bancarrota.
- Fracaso financiero: El fracaso financiero tiene dos fases bien diferenciadas. La primera, donde la empresa no puede hacer frente a los pagos estipulados en una fecha acordada, lo que genera un desequilibrio en la tesorería que puede afectar estructura económica de la empresa y tensiones de liquidez. Y la segunda es cuando esta deuda se prolonga en el tiempo generando un riesgo financiero aun mayor.

- Fracaso jurídico: Es la insolvencia definitiva, en la que se identifica a la empresa de manera judicial con un patrimonio neto negativo, lo que conlleva una sanción legal. La declaración de concurso en España puede ser presentada por el deudor, donde se debe justificar su endeudamiento y su insolvencia financiera. O por otra parte puede ser presentada por los acreedores 6 meses después de la fecha estipulada de la entrega del pago del crédito prestado al deudor.

En buena parte de la literatura encontrada, la definición de fracaso empresarial la encontramos como el fracaso jurídico descrito anteriormente, esto es así puesto que es una definición más objetiva que las otras. El problema de escoger esta definición es que puede haber empresas clasificadas como no fracasadas que estén pasando por una situación económica o financiera grave y que todavía no se hayan declarado en concurso de acreedores.

Este problema será solventado a medida que se hagan iteraciones del proceso de minería de datos. Esto hará que en una primera iteración escojamos todas las empresas declaradas como fracasadas y no fracasadas según la definición jurídica de fracaso, pero en posteriores iteraciones, gracias al conocimiento extraído de los modelos creados en la primeras iteraciones se irá corrigiendo la manera de escoger las muestras con las que trabajaremos. Así, se podrán descartar empresas activas que presenten en unas variables con un rendimiento bajo y que sean empresas que puedan presentarse como concurso de acreedores en un corto plazo de tiempo.

En esta primera iteración se escogerán por parte de las empresas fracasadas los años relativos anteriores al año del fracaso (t), ($t - 1$) y ($t - 2$). En el caso de las empresas que están en activo se escogerán los datos financieros relativos al ultimo año con datos disponibles (t), ($t - 1$) y ($t - 2$).

También será importante que estados fiscales vamos a escoger como empresas fracasadas y cuales no. SABI nos presentaba todas las posibilidades representadas en las figuras 4.1 y 4.2.

Como podemos observar en las figuras 4.1 y 4.2 lo más intuitivo es escoger a aquellas empresas que se encuentran inactivas. Pero esto, es un error puesto que estamos hablando de empresas que casi en todos los casos se habrán encontrado durante años en concurso de acreedores hasta su extinción total, lo que provocará que los patrones encontrados en nuestros modelos sean de empresas que ya se encuentran en una situación económica irreversible y no aquellas que es posible que en un corto plazo de tiempo, se encuentren en una situación económica difícil. Por ello, optaremos por las empresas



Figura 4.1: Etiquetas empresas Activas en SABI.



Figura 4.2: Etiquetas empresas Inactivas en SABI.

que siguen activas pero su situación fiscal es de concurso de acreedores, suspensión de pagos o quiebra.

Debido a la Ley Concursal ([BOE, 2014](https://www.boe.es/buscar/act.php?id=BOE-A-2003-13813))¹, toda empresa que no pueda hacer frente a sus impagos serán declaradas en concurso de acreedores, así el concepto de quiebra, suspensión de pagos y concurso de acreedores se unifica en concurso de acreedores. Comentar que esta ley entra en vigor con el fin de que las empresas que entren en concurso, que tienen proyectos rentables, pero que tienen problemas de carácter financiero y que no pueden hacer frente a la deuda, puedan seguir con su proyecto, los empleados no pierdan el puesto de trabajo y puedan seguir cubriendo la deuda adquirida con sus acreedores. Si consultamos el artículo 2.4 de dicha ley obtenemos: “Si la solicitud de declaración de concurso la presenta un acreedor, deberá fundarla en título por el cual se haya despachado ejecución o apremio sin que del embargo resultasen bienes libres bastantes para el pago, o en la existencia de alguno de los siguientes hechos:

- 1. El sobreseimiento general en el pago corriente de las obligaciones del deudor.
- 2. La existencia de embargos por ejecuciones pendientes que afecten de una manera general al patrimonio del deudor.

¹<https://www.boe.es/buscar/act.php?id=BOE-A-2003-13813>

- 3. El alzamiento o la liquidación apresurada o ruinosa de sus bienes por el deudor.
- 4. El incumplimiento generalizado de obligaciones de alguna de las clases siguientes: las de pago de obligaciones tributarias exigibles durante los tres meses anteriores a la solicitud de concurso; las de pago de cuotas de la Seguridad Social, y demás conceptos de recaudación conjunta durante el mismo período; las de pago de salarios e indemnizaciones y demás retribuciones derivadas de las relaciones de trabajo correspondientes a las tres últimas mensualidades.”

En el artículo 2.4 habla en los casos en los que el deudor puede declarar a su empresa en concurso de acreedores acreditando su endeudamiento y su estado de insolvencia para hacer frente a los pagos. Por otra parte, el concurso puede ser solicitado por alguno de sus acreedores dando prueba fehaciente de su estado de insolvencia.

“La unidad del procedimiento impone la de su presupuesto objetivo, identificado con la insolvencia, que se concibe como el estado patrimonial del deudor que no puede cumplir regularmente sus obligaciones. Pero ese concepto unitario es también flexible y opera de manera distinta según se trate de concurso necesario o voluntario. Los legitimados para solicitar el concurso del deudor (sus acreedores y, si se trata de una persona jurídica, quienes respondan personalmente de sus deudas) han de basarse en alguno de los hechos que como presuntos reveladores de la insolvencia enuncia la ley: desde la ejecución singular infructuosa hasta el sobreseimiento, general o sectorial, según afecte al conjunto de las obligaciones o a alguna de las clases que la ley considera especialmente sensibles en el pasivo del deudor, entre otros hechos tasados.

Incumbe al solicitante del concurso necesario la prueba de los hechos en que fundamentalmente su solicitud ; en todo caso, la declaración ha de hacerse con respeto de las garantías procesales del deudor, quien habrá de ser emplazado y podrá oponerse a la solicitud, basándose en la inexistencia del hecho en que ésta se fundamente o en la de su estado de insolvencia, incumbiéndole en este caso la prueba de su solvencia. Las garantías del deudor se complementan con la posibilidad de recurrir la declaración de concurso.”Ley Concursal ([BOE, 2014](#)).

Además, debemos de tener en cuenta que debido a la crisis, de carácter financiero, que comienza en 2008 y que se ve afectada por la caída de la empresa *Lehman Brothers*, provoca que el PIB (Producto Interior Bruto) tanto de la Unión Europea, Zona Euro y España como se puede ver en la figura 4.3 caiga en picado. No será hasta el 2012, cuando en España se empiecen a ver síntomas de recuperación económica.

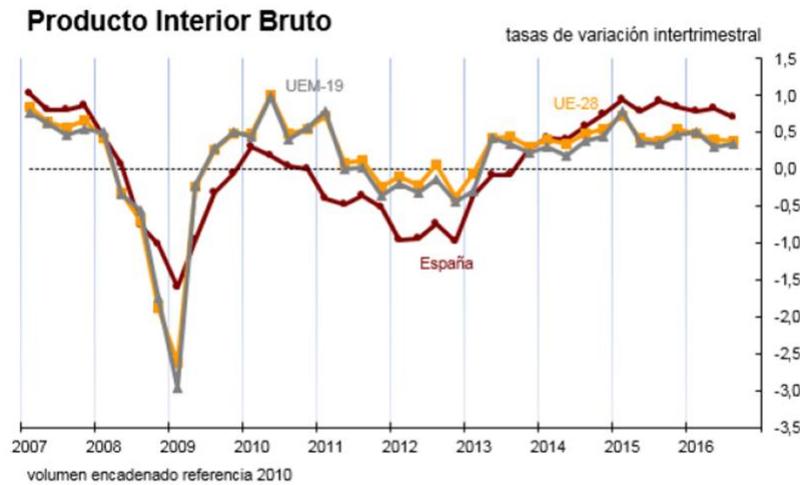


Figura 4.3: Evolución del PIB en España.



Figura 4.4: Evolución del porcentaje de parados en España.

Esto no solo afecta al PIB, si no que el desempleo crece de manera exponencial como podemos ver en la figura 4.4(Datos extraídos del INE²). Esto ocurre por el gran número de empresas fracasadas durante todos estos años como se ve en la figura 4.5, destacando el sector de la construcción. Además, hay que recordar que el grueso del empleo en España es el autoempleo, es decir, personas autónomas que trabajan por

²<https://www.ine.es/>

cuenta propia y que tuvieron que cesar su actividad económica por las causas anteriormente mencionadas o por no hacer frente a los pagos de sus acreedores o fruto de la crisis financiera en España.

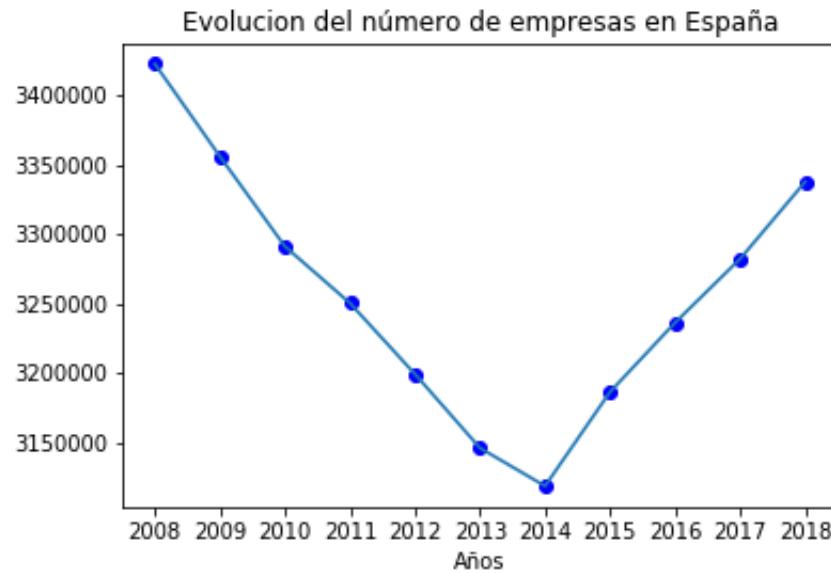


Figura 4.5: Evolución del número de empresas en España.

Puesto que en este TFG no tendremos variables desde un punto de vista macroeconómico, empresas con datos disponibles dentro de este periodo se descartarán por la posibilidad de que generen ruido en nuestros modelos de predicción.

Por último, y puesto que se va a hacer de manera reiterada menciones a diferentes variables se procede a dar una breve explicación:

- **Activos:** Un activo es un bien que tiene la empresa, que en un cierto plazo de tiempo puede convertirse en dinero o en otros medios líquidos equivalentes. Dentro de estos activos tenemos dos tipos. El activo fijo, que son aquellos activos utilizados en las empresas y no son adquiridos con fines de venta y los cuales no se pueden convertir en dinero en un corto plazo de tiempo. Por otra parte tenemos los activos líquidos que son aquellos que se esperan que sean utilizados en un periodo inferior al año como pueden ser las existencias.
- **Pasivos:** Los pasivos son las deudas que posee una empresa. Existen 2 tipos de pasivos, los fondos propios, que es el dinero que invierten los propietarios y les ha de ser devuelto. Y el pasivo exigible, que son las deudas que son reclamables a la empresa por alguien ajeno a ella. Esta última a su vez se divide en dos. El pasivo fijo, que es aquel que vence en un periodo superior a un año y por otra

parte el pasivo circulante o líquido que es aquel que vence en un plazo inferior a un año.

4.1.2. Estudio y compresión de los datos

Para comenzar se explicará como se han obtenido las muestras de empresas y que proceso seguimos para seleccionarla en la base de datos de SABI.

The screenshot shows the main interface of the SABI database. At the top, there's a header with the text "sabi 2.500.000 empresas españolas y 700.000 empresas portuguesas". Below the header, there are tabs for "Empresas", "Contactos", "Informes sectoriales", and "Noticias". A search bar is present with placeholder text "Nombre empresa o número BvD ID" and a magnifying glass icon. To the right of the search bar are links for "Alertas", "Personalizar", "Ayuda", "Contactarnos", and "Cerrar sesión". A sidebar on the left contains various filters like "Nombre empresa", "Número de identificación", "Estado", etc. The main content area has a sidebar menu on the right with sections such as "Datos financieros", "Empleados", "Ratios", "Leasing, Financiación, Subvenciones", "Incidencias", "Tipos de cuentas y disponibilidad", "Datos bursátiles", "Informes actualizados", "Datos personalizados", and "Todas las empresas". At the bottom of the page, there's a section titled "Su suscripción SABI le permite acceder a la siguiente información" with several items listed, each with a checkmark or a red circle. The footer indicates the last update was on "02/04/2019".

Figura 4.6: Pantalla inicial de la base de datos de SABI.

Como podemos ver en la figura 4.6 SABI presenta diferentes filtros relacionados con los datos financieros de las empresas. Pero, para esta primera iteración solo se tomará como filtro el estado fiscal de las empresas.

This screenshot shows the search results page in SABI. At the top, there's a section titled "ESTRATEGIA DE BÚSQUEDA" with a checkbox labeled "1. Estados España: Activa" which is checked. Below this, there's a "Búsqueda booleana" field containing the value "1". On the right side, there are buttons for "Guardar", "Imprimir", and "Borrar todas las etapas". The total number of results is displayed as "832.806". At the bottom, there's a button labeled "Ver lista de resultados".

Figura 4.7: Número de resultados una vez haber ejecutado los filtros en SABI

Después de seleccionar el filtro deseado, SABI nos puede mostrar los resultados de este, al aplicárselo a toda la base de datos 4.7. Posteriormente, si le damos a ver resultados nos muestra las empresas sin haber personalizado las columnas o variables que deseamos, para ello deberemos pinchar en la pestaña de columnas.

Como podemos ver en la parte izquierda de la figura 4.9 tenemos los diferentes datos fiscales a seleccionar y en la parte derecha el total seleccionado. En la imagen

	Nombre	Nombre	Localidad	Añadir
1.	<input checked="" type="checkbox"/> <input type="checkbox"/> MERCADONA SA	MERCADONA SA	TAVERNE ▲	
2.	<input checked="" type="checkbox"/> <input type="checkbox"/> REPSOL PETROLEO SA	REPSOL PETROLEO SA	MADRID	
3.	<input checked="" type="checkbox"/> <input type="checkbox"/> COMPAÑIA ESPAÑOLA DE PETROLEOS SAU	COMPAÑIA ESPAÑOLA DE PETROLEOS SAU	MADRID	
4.	<input checked="" type="checkbox"/> <input type="checkbox"/> REPSOL COMERCIAL DE PRODUCTOS PETROLIFEROS SA	REPSOL COMERCIAL DE PRODUCTOS PETROLIFEROS SA	MADRID	
5.	<input checked="" type="checkbox"/> <input type="checkbox"/> CEPSA TRADING SA	CEPSA TRADING SAU	MADRID	
6.	<input checked="" type="checkbox"/> <input type="checkbox"/> ENDESA ENERGIA SAU	ENDESA ENERGIA SAU	MADRID	
7.	<input checked="" type="checkbox"/> <input type="checkbox"/> EL CORTE INGLES SA	EL CORTE INGLES SA	MADRID	
8.	<input checked="" type="checkbox"/> <input type="checkbox"/> INDUSTRIA DE DISEÑO TEXTIL SA	INDUSTRIA DE DISEÑO TEXTIL SA	ARTEIXO	
9.	<input checked="" type="checkbox"/> <input type="checkbox"/> SEAT SA	SEAT SA	MARTOR	
10.	<input checked="" type="checkbox"/> <input type="checkbox"/> FORD ESPAÑA SL	FORD ESPAÑA SL	ALCOBEN	
11.	<input checked="" type="checkbox"/> <input type="checkbox"/> SOCIEDAD ESTATAL LOTERIAS Y APUESTAS DEL ESTADO SME SA	SOCIEDAD ESTATAL LOTERIAS Y APUESTAS DEL ESTADO SME SA	MADRID	
12.	<input checked="" type="checkbox"/> <input type="checkbox"/> TELEFONICA DE ESPAÑA SAU	TELEFONICA DE ESPAÑA SAU	MADRID	
13.	<input checked="" type="checkbox"/> <input type="checkbox"/> RENAULT ESPAÑA SA	RENAULT ESPAÑA SA	VALLADOC ▼	
14.	<input checked="" type="checkbox"/> <input type="checkbox"/> CENTROS COMERCIALES CARREFOUR SA			

Figura 4.8: Empresas clasificadas como activas

4.9 podemos ver como se pueden guardar un listado de datos fiscales sin hacer ningún tipo de selección, puesto que esto se hará en etapas posteriores.

Figura 4.9: Columnas de las empresas

Una vez que tenemos preparado el conjunto de datos que queremos descargar, le clickamos en exportar. Entonces, esta es la pestaña que aparece en la figura 4.10 en este caso elegimos la extensión .txt, y que tuviera como separador la coma para poder tratarlo como un archivo con la extensión csv. Con este número de columnas podríamos descargar hasta un conjunto de datos parcial de 2500 registros, pero la empresa SABI

nos bloqueo el acceso a toda la UCLM por “descarga masiva de datos”. Poniéndonos en contacto con la biblioteca, SABI accede a desbloquearnos el acceso a la descarga y nos informaran que solo podemos descargar paquetes de conjuntos de datos de 500 en 500 registros.

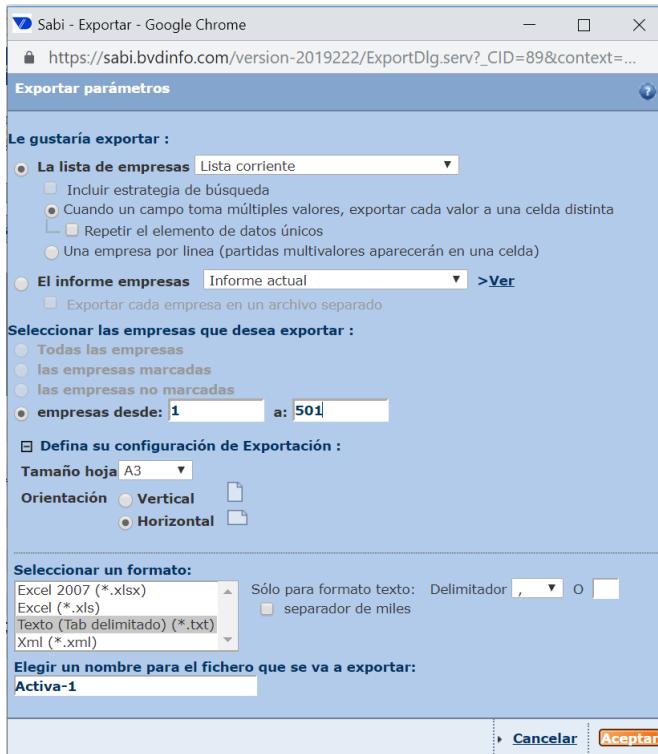


Figura 4.10: Exportacion del dataset

Para esta primera iteración, descargamos un total de 70,000 registros de empresas activas de más de las 800,000 que hay, y todas las que se encontraban en concurso de acreedores, que era un total de 5117.

El siguiente paso a la descarga de los datos es juntarlos en un único conjunto de datos para poder comenzar el análisis exploratorio de los mismos. En la descarga de los datos se puso para las empresas activas como nombre de archivo *Activas_numero de descarga.txt*. Posteriormente aplicamos el código adjuntado en el siguiente enlace³, en concreto en la clase *class_mount* para juntarlo en un solo archivo. Esto se realiza de manera análoga con la muestra de concursos. Para posteriormente, exportarlos como conjunto de datos final.

Es interesante poder estudiar si la actividad económica puede ser una característica

³https://github.com/JaimeTolosaDeLaFuente/TFG_JaimeTolosa/blob/master/Iteraci%C3%B3n%205/Cuaderno%20Jupyter/preprocesamiento.py

importante a la hora de generar el modelo final. En España, en concreto tenemos los códigos de CNAE, que contiene una lista completa de todas las actividades económicas. Son 21 actividades económicas diferentes, las cuales se dividen en 629 sub-actividades. El proceso para identificar el sector económico de cada empresa es el siguiente:

- Creación del fichero del CNAE: Copiamos y pegamos todos los códigos de la página del CNAE en un fichero de texto. Donde los nombres están separados por “.-”. Así, el fichero tendría la forma de la figura 4.11
- Generar DataFrame con los códigos del CNAE: Para generar un conjunto de datos a partir del fichero copiado y pegado de página web⁴. Se abre el txt con la función *open* de python. Entonces, linea a linea cogemos la primera parte con el separador ‘.-’ y si es una letra es el nuevo tipo de sector, si no se coge la letra de la sección y se pega el código y la subactividad.

```
A.- AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA
0111.- Cultivo de cereales (excepto arroz), leguminosas y semillas oleaginosas
0112.- Cultivo de arroz
0113.- Cultivo de hortalizas, raíces y tubérculos
0114.- Cultivo de caña de azúcar
0115.- Cultivo de tabaco
0116.- Cultivo de plantas para fibras textiles
0119.- Otros cultivos no perennes
0121.- Cultivo de la vid
0122.- Cultivo de frutos tropicales y subtropicales
0123.- Cultivo de cítricos
0124.- Cultivo de frutos con hueso y pepitas
0125.- Cultivo de otros árboles y arbustos frutales y frutos secos
0126.- Cultivo de frutos oleaginosos
0127.- Cultivo de plantas para bebidas
0128.- Cultivo de especias, plantas aromáticas, medicinales y farmacéuticas
0129.- Otros cultivos perennes
0130.- Propagación de plantas
0141.- Explotación de ganado bovino para la producción de leche
0142.- Explotación de otro ganado bovino y búfalos
0143.- Explotación de caballos y otros equinos
0144.- Explotación de camellos y otros camélidos
0145.- Explotación de ganado ovino y caprino
0146.- Explotación de ganado porcino
0147.- Avicultura
0149.- Otras explotaciones de ganado
0150.- Producción agrícola combinada con la producción ganadera
0161.- Actividades de apoyo a la agricultura
0162.- Actividades de apoyo a la ganadería
0163.- Actividades de preparación posterior a la cosecha
0164.- Tratamiento de semillas para reproducción
```

Figura 4.11: Fichero de las actividades económicas según el CNAE.

Ahora sí, se puede hacer es generar una nueva variable que a partir del conjunto de datos generado a partir del .txt y el código del CNAE proporcionado por el conjunto de datos de SABI. Esta, variable podría ser interesante, para su uso en el modelo final, para ello podemos realizar un histograma viendo cuáles son los sectores que más empresas fracasadas tienen y otro histograma donde veamos que sectores son aquellos que poseen más empresas activas.

⁴<https://www.cnae.com.es/lista-actividades.php>

Como podemos observar en la figura 4.12 y la figura 4.13, el sector con mayor número de empresas en concurso se dedican al sector de la construcción(todas las siglas se encuentran descritas en la tabla 4.1). Esto puede ser así puesto que es uno de los sectores con más empresas activas y por ello puede ser que en proporción también sea uno de los que más entren en el estado de concurso. Pero el hecho es, que la proporción se aumenta en el sector de la construcción, respecto de otras actividades que tienen un número de empresas activas alto.

Esto se puede explicar por la explosión de la burbuja inmobiliaria en España en 2009 y que hizo quebrar tanto empresas de construcción como inmobiliarias.

Por lo tanto, esta variable puede ser interesante a la hora de incluirla en el conjunto de datos final, puesto que junto con otras variables se puede convertir en una variable importante a la hora de discriminar el espacio de clasificación. Este tipo de variables también ha sido usada en otros estudios como el de (Alfaro, 2008).



Figura 4.12: Frecuencia de empresas fracasadas según el tipo de actividad.

Letra de la actividad	Actividad
A	Agricultura, ganadería, silvicultura y pesca
B	Industrias extractivas
C	Industria manufacturera
D	Suministro de energía eléctrica, gas, vapor y aire acondicionado
F	Construcción
G	Comercio al por mayor y al por menor
H	Transporte y almacenamiento
I	Hostelería
J	Información y comunicaciones
K	Actividades financieras y de seguros.
L	Actividades Inmobiliarias
M	Actividades profesionales, científicas y técnicas
N	Actividades administrativas y servicios auxiliares.
O	Administración pública y defensa; seguridad social obligatoria.
P	Educación
Q	Actividades sanitarias y de servicios sociales
R	Actividades artísticas, recreativas y de entretenimiento
S	Otros servicios.
T	Actividades de los hogares empleadores de personal doméstico
U	Actividades de organizaciones y organismos extraterritoriales.

Tabla 4.1: Actividades por letra CNAE 2009.

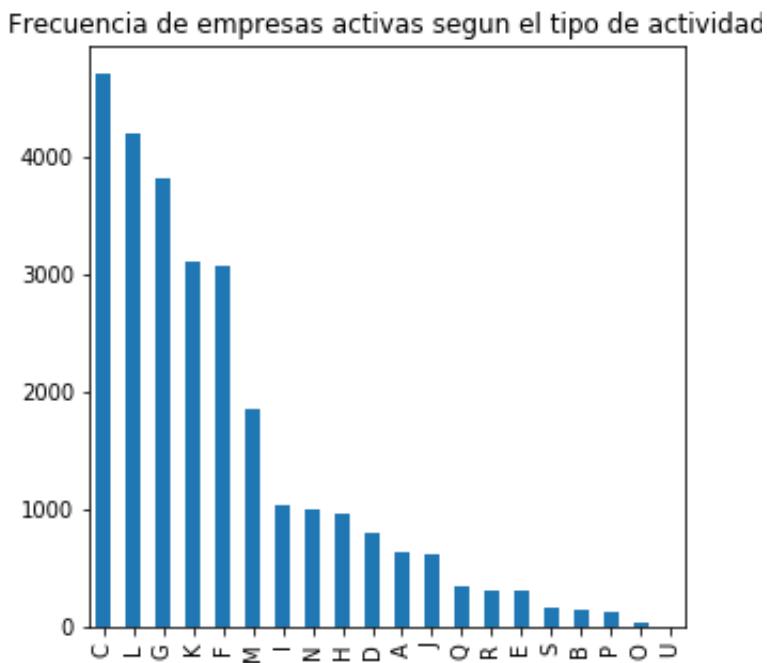


Figura 4.13: Frecuencia de empresas activas según el tipo de actividad.

Preprocesamiento de los datos

En este punto del preprocesado es importante en tanto y en que forma se distribuyen los valores nulos a lo largo de nuestro conjunto de datos. Para ello Pandas nos aporta una infinidad de formas de hacerlo.

Existen métodos en la librería Pandas para ver cuantos valores nulos tenemos por columna. El problema de hacerlo de esa manera es que se tiene que mirar cada columna, una a una, para saber que cantidad de nulos son los que tenemos en cada columna.

Puesto que estas variables contienen una gran cantidad de nulos, una forma de encontrarlos es simplemente ver si la cantidad de nulos que hay supera la mitad del tamaño de instancias del conjunto de datos. Este conjunto de variables con valores nulo, indica que hay tipos de datos financieros están en desuso o han pasado a denominarse de otra manera y por ello en los últimos años las empresas no las han declarado.

En estos momentos, el conjunto de datos esta preparado para empezar a empezar a imputar los valores nulos que contengan. El problema principal es que aún habiendo eliminado bastantes columnas que contenían una cantidad de nulos muy alta aún quedan otras muchas variables que contienen alrededor de un $\frac{1}{3}$ de las instancias con

valores nulos. Y por tanto, la media no parece una buena forma de imputar esos valores nulos, puesto que puedes hacer variables que las variables contengan una varianza muy baja y que no aporten información alguna a los modelos.

Una opción es eliminar todas aquellas variables para las cuales el número de nulos supere $\frac{1}{10}$ parte del conjunto total del instancias, pero para estas primeras opciones se quiere conservar el mayor número el número de instancias y variables posible. En posteriores iteraciones se irá refinando este tipo parámetros haciendo que sean cada vez más restrictivos y obtener un conjunto final con unos datos de calidad.

Otra manera sería imputar los valores nulos de la manera más efectiva posible. Como comentamos al principio del documento el conjunto de datos consta de los datos financieros de las empresas durante 3 ejercicios diferentes. Por ello se generan submuestas donde el conjunto de datos no contuviera ninguna instancia nula en ese trío de columnas de cada variable relativa al año (t) , $(t - 1)$ y $(t - 2)$.

De esa manera se puede crear un modelo para predecir el valor de cada variable a partir de los otros dos años relativos restantes. Y en el caso de faltar dos años relativos se predice cada uno de ellos a partir del existente.

Este modelo de regresión para los valores nulos lo que hace es una vez entrenado todos los modelos para las variables, recorre el conjunto de datos imputando los valores nulos que vaya encontrando.

Ahora, se debe decidir que modelo de regresión sera elegido para utilizar como imputador. Para hacer la comparativa entre los modelos de regresión, se usa el conjunto de validación donde se obtienen todas aquellas filas, que no contuvieran valores nulos y se predecían con el imputador.

Una vez imputado, se procede a hacer la Suma de las Diferencias de los Cuadrados (SSD) entre los valores predecidos y el conjunto de validación, pasandolo antes por un escalado para que todas las variables contuviera la misma magnitud.

Para decidir cual sería nuestro imputador probamos 3 modelos sencillos que fueran rápidos en entrenamiento y en inferencia. En primer lugar, el Árbol de Regresión⁵ fue el mejor de todos los modelos probados como podemos ver en la tabla 4.2. La

⁵<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor>

Regresión Lineal⁶ en cambio, fue el peor de los modelos, esto es debido a la simpleza del modelo. Y por último, K Vecinos más cercanos que se comportó bien para la mayoría de las predicciones, pero los Árboles de Regresión son los que mejor comportamiento mostraron.

	Árbol de Regresión	Regresión Lineal	K Vecinos más cercanos
SSD	2757786.77	$2.58 \cdot 10^{19}$	111922156.03

Tabla 4.2: Suma de la diferencias de los cuadrados según el modelo de regresión.

Una vez se usa el imputador de valores nulos, la mayoría de valores nulos desaparecen como vemos en la tabla 4.3 que son el ejemplo de algunas de las variables. Como el número de instancias que contienen valores nulos es bajo, los valores restantes se imputarán por la media.

VARIABLES	NÚMERO DE NULOS ANTES DE IMPUTAR	NÚMERO DE NULOS DESPUES DE IMPUTAR
INGRESOS DE EXPLOTACIÓN EUR -2	2379	186
RENTABILIDAD ECONOMICA (%) % AÑO -2	1693	0
RENTABILIDAD FINANCIERA (%) % AÑO -2	1702	7

Tabla 4.3: Reducción del número de nulos tras aplicar el imputador.

Ahora, que nuestro conjunto de datos no contiene ningún valor nulo, debemos eliminar aquellas variables del conjunto de datos que identifican a cada instancia de manera única. Y que no aporta ningún tipo de conocimiento al problema. Las variables eliminadas serán Nombre, Código NIF, Código numérico del CNAE, Comunidad Autónoma, Localidad y Última fecha disponible.

Ahora, debemos convertir las variables cualitativas en variables cuantitativas, se ha optado por esta opción puesto que hay muchos algoritmos como por ejemplo las RNNs, XGBoost, etc. que tan solo trabajan con variables numéricas. A este tipo de variables, se les denominan *dummy variable*, una manera de transformar estas variables cualitativas en variables cuantitativas es eliminando esa variable, y cada uno de los valores cualitativos que pueda tener esa variable transformarlos en una nueva variable y adjudicarle un 1 o un 0 según si ese registro tenía ese valor cualitativo o no.

Este tipo de variables se pueden crear de manera automática con la librería Pandas que tiene el método *get_dummies*⁷, donde al pasarle una variable con valores cualitati-

⁶https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

⁷https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html

vos, te devuelve un *DataFrame* con las nuevas variables creadas a partir de la variable cualitativa.

4.1.3. Selección y transformación de las variables

En esta primera iteración no se hizo ningún tipo de selección o transformación de los datos, puesto que se quería hacer una iteración rápida, para observar que resultados obteníamos sin hacer ningún tipo de tratamiento de selección o transformación a los datos.

4.1.4. Modelado

Una vez llegado a este punto, en esta primera iteración se quisieron probar modelos muy sencillos como podía ser un Árbol de Decisión.

Al hacer una validación cruzada con la librería de ([Scikit-Learn, 2019](#)) con un Árbol de Decisión⁸ de profundidad 1 como máximo, que en principio son Árboles de Decisión que tienen un error debido al sesgo alto y usando la métrica $F_1 score$, obteníamos un 93.98 % de acierto, lo cual es un porcentaje muy alto para un Árbol de Decisión tan sencillo.

En el caso de usar un ensemble, que son modelos mucho más potentes que los Árboles de Decisión clásicos, como un Random Forest sin límite de profundidad obtenemos un 99.3 % de acierto con la métrica $F_1 score$.

El motivo por el cual se dieron estos resultados podría ser o que el problema es demasiado fácil o se hizo una selección de la muestra muy mala, de tal manera que se han obtenido muestras de empresas activas muy diferentes a las muestra de empresas en concurso.

4.1.5. Puesta en producción

Debido a los “malos” resultados era necesario volver a hacer otra iteración desde la fase de estudio y comprensión de los datos, para ver si eramos capaces de detectar, porque estos datos habían tan sido tan fáciles de clasificar.

⁸<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

4.2. Segunda iteración

4.2.1. Estudio y compresión de los datos.

En esta segunda iteración, y viendo los resultados que obtiene ([Altman, 1968](#)), con un modelo matemático como el análisis discriminante multivariable del 95 % de precisión, prediciendo con los datos financieros del año en el que la empresa cae en quiebra, se llega a la conclusión que ese sería el motivo por el cual nuestros datos eran muy fáciles.

Entonces, una solución al problema es hacer una selección de la muestra con los años $(t - 2)$, $(t - 3)$ y $(t - 4)$ relativos al año de quiebra (t) de la empresa. De esta manera identificaríamos, los primeros síntomas del fracaso empresarial y no un estado de no retorno del fracaso empresarial. Para ello se vuelve a realizar otra vez la descarga del nuevo conjunto de datos con SABI.

Esta empresa vuelve a denegar el acceso a su base de datos, por otra “descarga masiva de datos”. Mencionar, que esto fue un problema durante todo el proyecto puesto que una de las tareas en la que más esfuerzo se invirtió en tiempo fue la selección de la muestra y la descarga de datos ya que la herramienta era muy poco flexible a la hora de poder elegir muestras, P.e. algo tan sencillo como hacer un muestreo aleatorio de la base de datos, no se podía hacer.

Además no se podía descargar el conjunto de datos en una sola descarga, si no que nos debíamos descargar nuestra muestra de 500 en 500 registros, como se ha comentado en el punto anterior. Lo que hace que cada vez que queremos obtener un nuevo conjunto de datos se debe de realizar en algunas ocasiones más de 100 descargas de manera manual.

Para seleccionar que año queremos descargar de las variables. Simplemente, una vez que se selecciona la variables que se quiere utilizar para nuestra descarga, SABI nos muestra un dialogo, respecto a que año relativo queremos los datos de esa variable. Una vez obtenemos los datos de las empresas en concurso, se procede con la siguiente fase de la iteración.

4.2.2. Análisis de los datos, selección y transformación

Esta iteración es análoga a la explicada en la iteración anterior y no se introduce ningún cambio.

4.2.3. Modelado

El modelado que hicimos fue semejante que a la anterior iteración. En la prueba del Árbol de Decisión con una profundidad maxima de 1 obtuvimos resultados semejantes a la iteración anterior. En el caso del Random Forest⁹ ocurre lo mismo, la clasificación es demasiado buena, y ahora si debemos indagar en el porque ocurre esto.

Una manera muy sencilla de ver que esta ocurriendo, es visualizar un Árbol de Decisión, esta vez con una profundidad máxima de 3, para observar cuales son las separaciones del espacio de clasificación y porque se están dando estos resultados tan buenos.

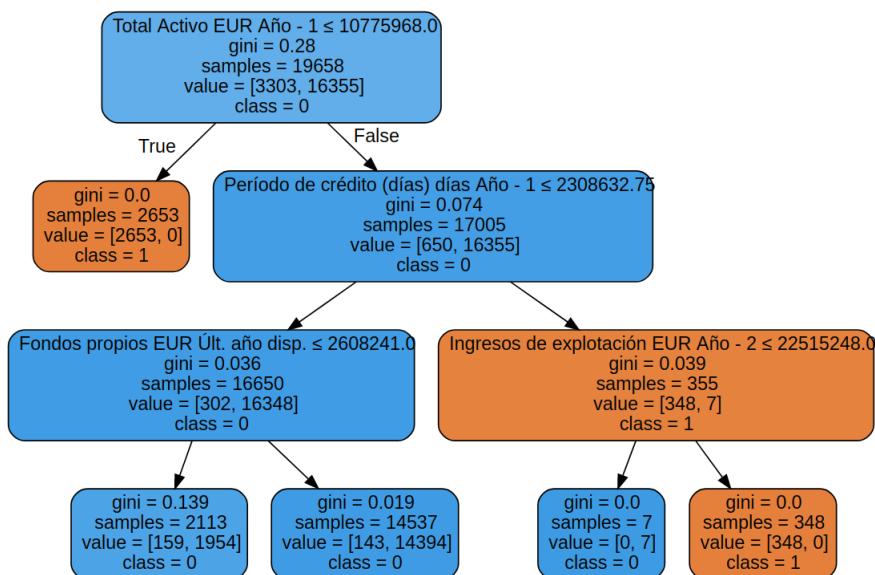


Figura 4.14: Visualización del del Árbol de Decisión

Como podemos observar en la figura 4.14 el gran problema es el activo total, y es que solo con esta variable es capaz de hacer una separación del espacio de clasificación con un acierto igual a $\frac{17005}{19658}$. Esto hace ver que el problema ha sido en la selección de la muestra puesto que hemos escogido empresas que no son similares en valores absolutos.

⁹<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

4.3. Tercera iteración

4.3.1. Estudio y compresión de los datos.

En esta tercera iteración, vamos observar como se ordenan las instancias en la base de datos de SABI, para poder averiguar cual ha sido el error a la hora de hacer la selección de la muestra. Como podemos observar en la figura 4.15, parece ser que la ordenación de la base de datos es de mayor a menor de la variable “Ingresos de explotación mil EUR Últ. año disp.”. Recordamos que en las anteriores iteraciones nosotros escogimos las 70.000 primeras sin ningún tipo de filtro, seleccionando aquellas que más ingresaron en su ultimo año disponible por su ejercicio. Algo que esta íntimamente relacionado con el tamaño de la empresa, sesgando totalmente los modelos generados.

	Nombre	E 2009	X	Ultimo año disponible	X	Ingresos de explotación mil EUR Últ. año disp.	X	Ingresos de explotación mil EUR Año - 1
1.	<input checked="" type="checkbox"/> REGULADORA DE COMPRAS DEL MEDITERRANEO SA			31/12/2017		1.210.992		1.151.039
2.	<input checked="" type="checkbox"/> FORTIA ENERGIA SL			31/12/2017		640.875		521.915
3.	<input checked="" type="checkbox"/> GM FUEL SERVICE SL			31/12/2017		638.104		320.815
4.	<input checked="" type="checkbox"/> HIPER USERA SL			31/12/2016		329.182		334.785
5.	<input checked="" type="checkbox"/> UNICA GROUP S.C. ANDALUZA			31/08/2017		317.144		258.611
6.	<input checked="" type="checkbox"/> SISTEMA D'EMERGENCIES MEDIQUEES SA			31/12/2017		295.951		209.606
7.	<input checked="" type="checkbox"/> GNERA ENERGIA Y TECNOLOGIA SL			31/12/2017		271.602		231.187
8.	<input checked="" type="checkbox"/> SANLUCAR FRUIT SL			30/06/2017		236.163		198.147
9.	<input checked="" type="checkbox"/> ACEITES ABRIL SL			31/12/2017		233.964		161.353
10.	<input checked="" type="checkbox"/> CEPSA GAS Y ELECTRICIDAD SA,			31/12/2017		183.405		127.185
11.	<input checked="" type="checkbox"/> ESERGUI DISTESER SOCIEDAD LIMITADA.			31/12/2017		175.340		72.467
12.	<input checked="" type="checkbox"/> BIOTERUEL SL			31/12/2017		172.966		29.583
13.	<input checked="" type="checkbox"/> IBERLECHE S.L.			31/12/2017		171.042		140.012
14.	<input checked="" type="checkbox"/> AGENCIA PUBLICA EMPRESARIAL SANITARIA COSTA DEL SOL			31/12/2017		169.813		167.761
15.	<input checked="" type="checkbox"/> SERHS DISTRIBUCIO I LOGISTICA SL			31/08/2018		169.406		183.267
16.	<input checked="" type="checkbox"/> ATUNES Y LOMOS SL			31/12/2017		165.997		122.038
17.	<input checked="" type="checkbox"/> S.C. AND ALMAZARAS DE LA SUBBETICA			30/09/2017		164.220		103.037
18.	<input checked="" type="checkbox"/> CENTRO PORTUARIO DE EMPLEO DE VALENCIA ETT SOCIEDAD ANONIMA.			31/12/2017		161.133		155.565
19.	<input checked="" type="checkbox"/> COMERCIAL PERNAS SL			31/12/2017		154.595		135.512
20.	<input checked="" type="checkbox"/> INTEROLEO PICUAL JAEN SOCIEDAD ANONIMA			31/12/2017		151.052		136.703
21.	<input checked="" type="checkbox"/> BORGES BRANDED FOODS SL			31/05/2018		150.602		149.184
22.	<input checked="" type="checkbox"/> CENTRO FARMACEUTICO DEL NORTE SOCIEDAD ANONIMA			31/12/2017		142.273		133.415
23.	<input checked="" type="checkbox"/> LIEBHERR IBERICA, S.L.			31/12/2017		138.302		101.181
24.	<input checked="" type="checkbox"/> JOYERIA FINA SL			31/12/2017		136.375		98.379
25.	<input checked="" type="checkbox"/> GRUPO REGIONAL DE COOPERATIVAS PLATANERAS DEL ARCHIPIELAGO C...			31/12/2017		129.893		119.744

Figura 4.15: Ordenación de la base de datos de SABI

Ahora, era necesario conocer de que manera podíamos medir el tamaño de una empresa. Según (Alfaro, 2008), el tamaño de la empresa se puede medir como el logaritmo natural del Activo Total. En este momento, se procede a hacer un estudio de como se distribuye esta variable en el conjunto de datos de las empresas declaradas en concurso para tener una muestra de empresas tanto activas como en concurso semejante.

Una vez que hemos aplicado el logaritmo natural a los 3 ejercicios sobre la variable del “Activo Total”, podemos hacer una visualización para comprobar si esta sufre

grandes variaciones de unos ejercicios a otros. Como podemos en la figura 4.16, esta no sufre variaciones durante los diferentes ejercicios lo que hace que la visualización se vea como una linea recta en el espacio tridimensional que pasa de un extremo al otro.

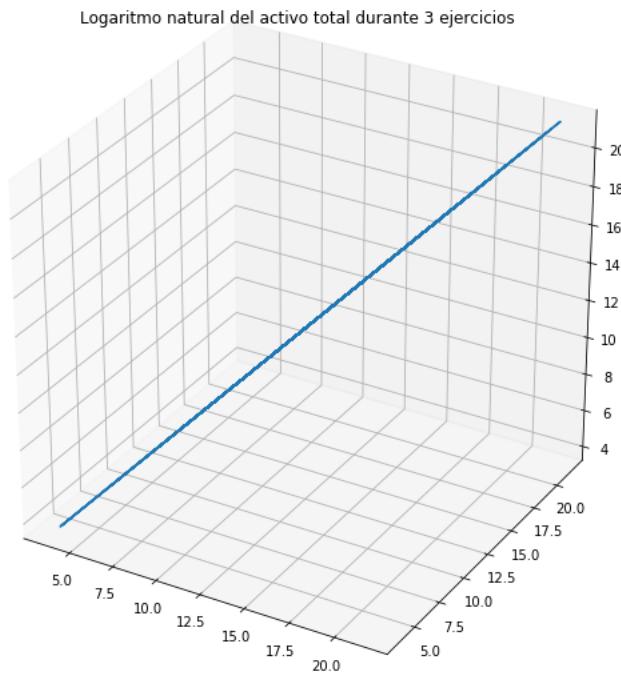


Figura 4.16: Logaritmo del activo total

En la figura 4.17, podemos observar como la mayoría de empresas declaradas en concurso, tienen un activo total muy bajo y que son muy pocas las empresas que se declaran en concurso cuando tienen un activo total alto. Esto explicaría porque en la anterior iteración el activo total era una variable para discriminar el estado de una empresa, puesto que los ingresos percibidos en un ejercicio guarda una correlación moderada respecto del activo total, en concreto del 0.61.

Posteriormente, se realiza un diagrama de cajas como podemos ver en la figura 4.18 del logaritmo del activo total para ver donde se encuentran los valores atípicos de la distribución. En la tabla 4.4 tenemos los resultados de este diagrama de caja, donde si elevamos el número e al bigote superior y al bigote inferior, obtenemos el límite inferior y superior a partir del cual la empresa tendría un tamaño tan bajo o alto,

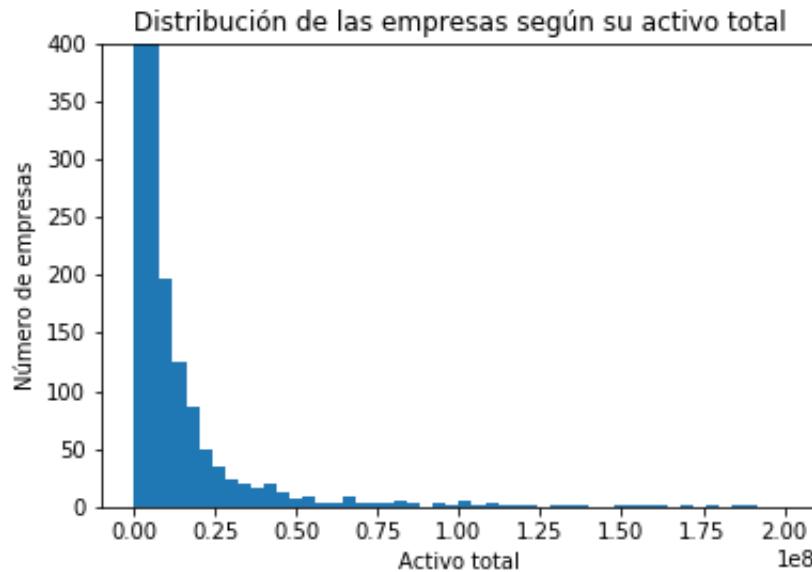


Figura 4.17: Distribución del activo total de empresas en concurso

que se considerarían *outliers*. Este límite inferior y superior sería entonces 10,777.118 y 126,631,582.071.

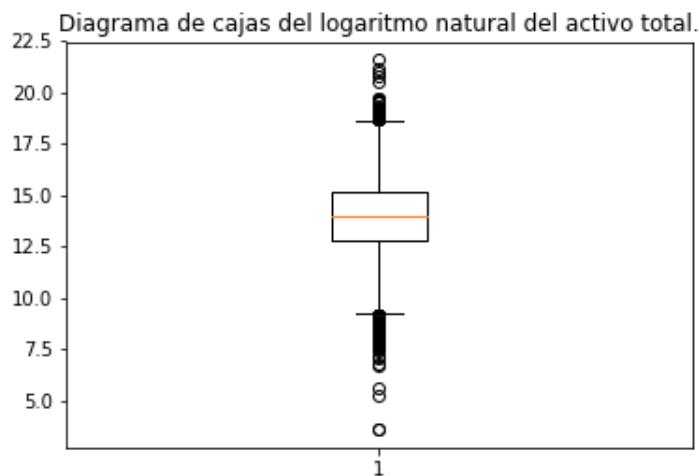


Figura 4.18: Diagrama de cajas del tamaño de las empresas.

Ahora añadiendo un filtro a SABI, como en la figura 4.19, podemos obtener una nueva muestra de datos con la que trabajar, en la que se supone estaremos trabajando con una muestra donde las empresas son similares y podremos obtener conclusiones sobre el fracaso empresarial.

Una vez aplicado el filtro obtenemos una muestra de un total de 685,018 empresas. El problema, en este momento es que descargarse las 685,018 en conjuntos de datos de

Percentil 25 %	Percentil 75 %	Media	Bigote superior	Bigote inferior
12.799	15.142	13.950	18.656	9.285

Tabla 4.4: Datos del diagrama de cajas del tamaño de las empresas

The screenshot shows the SABI platform interface. At the top, there's a header with the logo 'sabi' and the text '2.600.000 Spanish and 800.000 Portuguese companies'. Below the header, there's a navigation bar with tabs: 'Empresas' (selected), 'Contactos', 'Informes sectoriales', and 'Noticias'. A search bar follows, containing the placeholder 'Nombre empresa o número BvD ID' and a search icon. Below the search bar is a breadcrumb trail: 'Inicio > Buscar por perfil financiero & empleados'. The main content area has several filter sections: 'Partidas principales' (selected), 'Global', 'Formato detallado España', 'Ratios formato Global', 'Ratios españoles', and 'Ratios portugueses'. Under 'Seleccionar una variable', 'Total Activo' is chosen. Under 'Seleccionar una divisa', 'mil EUR' is chosen. Under 'Seleccionar periodos', 'Años absolutos' is selected, and a dropdown menu shows 'Datos anuales' with various year ranges checked (Ult. año disp., Ult. año disp. -1, Ult. año disp. -2, etc.). There's also a section for 'Criterio para validar' with the option 'al menos uno de los años seleccionados'. At the bottom, there are buttons for 'Buscar en': 'valores' (selected), 'Valor' (radio button), 'Mínimo' (text input: 10777 mil EUR), 'Máximo' (text input: 126631582 mil EUR), and 'Top/Cuartil' (dropdown). A note at the bottom states: 'Búsquedas de empresas con varios tipos de cuentas toman en cuenta las cuentas no consolidadas, excluyendo cuentas no recientes cuando otras cuentas existen [Modificar filtro]'.

Figura 4.19: Como hacer filtros en la plataforma SABI.

500 en 500 instancias es algo impracticable manualmente, ni tampoco podemos escoger los 50,000 primeros porque entonces se incide otra vez en el mismo error de las iteraciones anteriores. Además, tampoco podemos pedir que SABI te de un subconjunto de la muestra aleatorizada, porque no proporciona esa opción.

La manera de poder minimizar la muestra es seguir aplicando filtros a este conjunto de datos para obtener un subconjunto de datos más pequeño. Una manera de poder seguir disminuyendo el número de instancias sin sesgar el conjunto de datos a base de filtros, puede ser filtrando por aquellas empresas que tengan cierto conjunto de variables sin nulos. Pero esto último, también puede ser un problema pues como apunta ([Almeida, 2009](#)), las empresas con mayor capitalización tienden a aportar mayor información financiera y no financiera, puesto que esto, es una manera que aportar cierta confianza a los inversores. Con lo cual, volveríamos al problema del principio, tendríamos un conjunto de datos sesgado al tamaño de la empresa.

Por tanto, aplicaremos algunos filtros a partir de los datos obtenidos del diagrama caja y bigotes realizado para algunas variables predictoras, para que la muestra de empresas de concursos sea lo más semejante en cuanto a cantidad de ingresos, activo

total, etc. Así los discriminadores no harán separaciones en el espacio de clasificación por variables en tamaño absoluto, si no porque realmente una variable haga una correcta separación de la muestra, que sea extrapolable a la población. Por otra parte, también se aplicarán sin excedernos, filtros para escoger aquellas empresas que no tengan valores nulos.

En la tabla 4.5, se pueden ver los resultados tras aplicar el diagrama de cajas en algunas variables que pueden sesgar el problema. Para el caso, de datos financieros que sean negativos en el diagrama de cajas de las empresas en concurso, se escogerán empresas desde el 0 hasta el bigote superior de ese dato y al menos solo uno de los 3 años es suficiente, para asegurarnos de alguna manera que esas empresas gozan de una buena salud económica.

	Percentil 25 %	Percentil 75 %	Media	Bigote superior	Bigote inferior
Tamaño	12.8	15.1	13.9	18.6	9.3
Ingr. explotación	248725.5	2126511.0	3095453.2	4943189.2	-2567952.8
EBIT	-39776.0	65020.0	-93215.852	222214.0	-196970.0
EBITDA	-21057.5	102923.75	-19963.275	288895.625	-207029.375
Endeudamiento (%)	66.760	98.139	257.925	145.2099	19.690
Ingr. financieros	0.0	2135.5	75716.258	5338.75	-3203.25
Fondo de maniobra	7619.0	582207.0	954135.938	1444089.0	-854263.0

Tabla 4.5: Datos del diagrama de cajas de diferentes variables

La tabla 4.5, revela datos muy interesantes de la muestra de empresas en concurso, ya que se encuentran con resultados del ejercicio, aun ignorando intereses y costes financieros, en negativo. Además esto le puede llevar a un gran endeudamiento puesto que sin tener en cuenta este costo financiero han perdido dinero en el transcurso del ejercicio y serán incapaces de seguir pagando a sus deudores, lo que lleva a que el pasivo total crezca.

Lo explicado anteriormente se refleja de hecho en el endeudamiento, esta situación de pérdida de liquidez año tras año, hace que la deuda total crezca frente su activo, de hecho entre el percentil 25 % y el percentil 75 %, el endeudamiento crece hasta casi el 100 %, lo que quiere decir que la suma total de la deuda, es prácticamente el valor de su activo total. Recordar que estos datos son de los dos años anteriores al ejercicio donde se declaran en concurso, lo que quiere decir que los síntomas del fracaso aparecen tiempo antes de que este se declare.

		<input type="button" value="Guardar"/> <input type="button" value="Imprimir"/> <input type="button" value="Borrar todas las etapas"/>
	ESTRATEGIA DE BÚSQUEDA	
		<input type="button" value="Resultado etapa"/> <input type="button" value="Resultado búsqueda"/>
	1. Total Activo (mil EUR): Últ. año disponible, Último año -1, Último año -2, para al menos uno de los períodos seleccionados, min=10.777, max=126.631,582	1.562.147 1.562.147
	2. Estados España: Activa	834.827 812.490
	3. Ingresos de explotación: Todas las empresas con un valor conocido, Últ. año disponible	1.357.531 783.238
	4. Resultado del Ejercicio (mil EUR): Últ. año disponible, Último año -1, Último año -2, para al menos uno de los períodos seleccionados, min=0, max=129.7245	1.086.671 619.496
	5. Ingresos de explotación: Todas las empresas con un valor conocido, Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados	1.078.633 543.641
	6. Total activo: Todas las empresas con un valor conocido, Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados	1.344.417 543.641
	7. Resultado del Ejercicio: Todas las empresas con un valor conocido, Últ. año disponible, Último año -1, Último año -2, para al menos uno de los períodos seleccionados	1.595.381 543.641
	8. Fondo de Maniobra (€): Todas las empresas con un valor conocido, Últ. año disponible, Último año -1, Último año -2, para al menos uno de los períodos seleccionados	1.338.229 537.512
	9. Endeudamiento (%): Todas las empresas con un valor conocido, Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados	1.344.273 537.508
	10. Fondos propios: Todas las empresas con un valor conocido, Últ. año disponible, Último año -1, Último año -2, para al menos uno de los períodos seleccionados	1.637.059 537.508
	11. Ingresos financieros: Todas las empresas con un valor conocido, Últ. año disponible, Último año -1, Último año -2, para al menos uno de los períodos seleccionados	1.542.755 537.508
	12. Deudas financieras: Todas las empresas con un valor conocido, Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados	251.556 138.135
	13. Ingresos de explotación (mil EUR): Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados, min=0, max=943.18925	1.016.326 127.937
	14. Activo circulante (mil EUR): Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados, min=0, max=4.742.475	1.270.156 126.779
	15. EBIT (mil EUR): Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados, min=-196.97, max=222.214	1.089.191 118.285
	16. EBITDA (mil EUR): Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados, min=-207.029, max=288.895	1.100.664 116.207
	17. Endeudamiento (%): Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados, min=0, max=145	1.147.223 111.465
	18. Ingresos financieros (mil EUR): Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados, min=0, max=5.338	975.041 95.380
	19. Fondos propios (mil EUR): Últ. año disponible, Último año -1, Último año -2, para todos los períodos seleccionados, min=-854.263, max=1.444.089	92.744 1.195.945
	Búsqueda booleana	TOTAL : 92.744
	<input type="button" value="Actualizar"/>	
	<input type="button" value="Ver lista de resultados"/>	

Figura 4.20: Filtros de la muestra

Además, los ingresos financieros son muy bajos. Esto se debe a que su falta de liquidez hace que la gente no confíe en este tipo de empresas para financiarlas durante otro ejercicio, lo que hace que la empresa no pueda seguir sustentándose.

En la figura 4.20, vemos como hemos tenido que aplicar una gran cantidad de filtros de las variables estudiadas con el diagrama de cajas, para ir reduciendo poco a poco la muestra. Aun así, la cantidad de registros es bastante grande, ya que habría que hacer unas 186 descargas de 500 en 500 para obtener toda la muestra. Por temor a volver a sesgar la muestra a base de filtros se decide obtener esta muestra entera, aunque el coste de tiempo sea grande.

De hecho, si el coste en tiempo iba a ser grande, volvemos a tener problemas con la plataforma de SABI. Esta, vuelve a cancelar las descargas de datos financieros cuando el conjunto de datos ya ronda los 45,000 registros. En vista de esto, lo que se hace es pedir a un usuario de otra universidad para que colaboren en la descarga.

4.3.2. Análisis de los datos, selección y transformación de las variables

En primer lugar es necesario, hacer una submuestra de la muestra descargada ya que el número de instancias de empresas en concurso son 4800, mientras que en el caso de las empresas activas son 92,000 aproximadamente, esto es un problema tan desbalanceado que es muy difícil que los modelos obtengan patrones reales del problema, ya que con predecir todo como activo obtendría un 94 % de acierto y esto es algo con lo que lidiaremos posteriormente en la parte del modelado.

Para hacer la selección de la submuestra se hace de forma aleatoria con la función *shuffle* de la librería ([Numpy, 2019](#))¹⁰, donde de la clase mayoritaria escogemos 24,000 empresas escogidas de manera aleatoria. Así el problema desbalanceado será de una empresa en concurso contra 5 activas, aproximadamente.

Una vez aplicamos el mismo preprocessado para los valores nulos como en anteriores iteraciones, debemos empezar a disminuir la variabilidad de nuestro problema.

Dos variables que vamos a considerar en nuestro problema, eran las actividades económicas descritas por el CNAE y la forma jurídica de las empresas, ya que podrían ser interesantes en este tipo de problemas. Estas variables las convertimos en *dummies*

¹⁰<https://www.numpy.org/>

variables, para que pasaran de ser variables categóricas a numéricas, y puesto que contenían muchos valores categóricos aumentaban de una manera considerable el número de variables por ello sería interesante que aquellas que aparezcan con menor frecuencia agruparlas en una sola variable.

Para ello, en el caso de las actividades económicas están distribuidas como se ve en la figura 4.21, donde las actividades G,C,F y M son aquellas que destacan donde hay un número de empresas considerable y destacan sobre las demás actividades. Todas las restantes, pasan a formar parte de una misma variable que denominaremos “Otras actividades”.



Figura 4.21: Número de empresas según su tipo de actividad.

Para la variable “Forma Jurídica” su histograma tiene la forma de la figura 4.22, donde las únicas dos que son observables por su frecuencia son las sociedades limitadas y las sociedades anónimas. Por este motivo, será mejor ver como se distribuyen numéricamente como tenemos en la tabla 4.6, donde parece evidente que la sociedades limitadas y sociedades anónimas, serán aquellas que se mantengan como variables y todas las demás se agrupen como otra variable denominada “Otras Formas Jurídicas”.

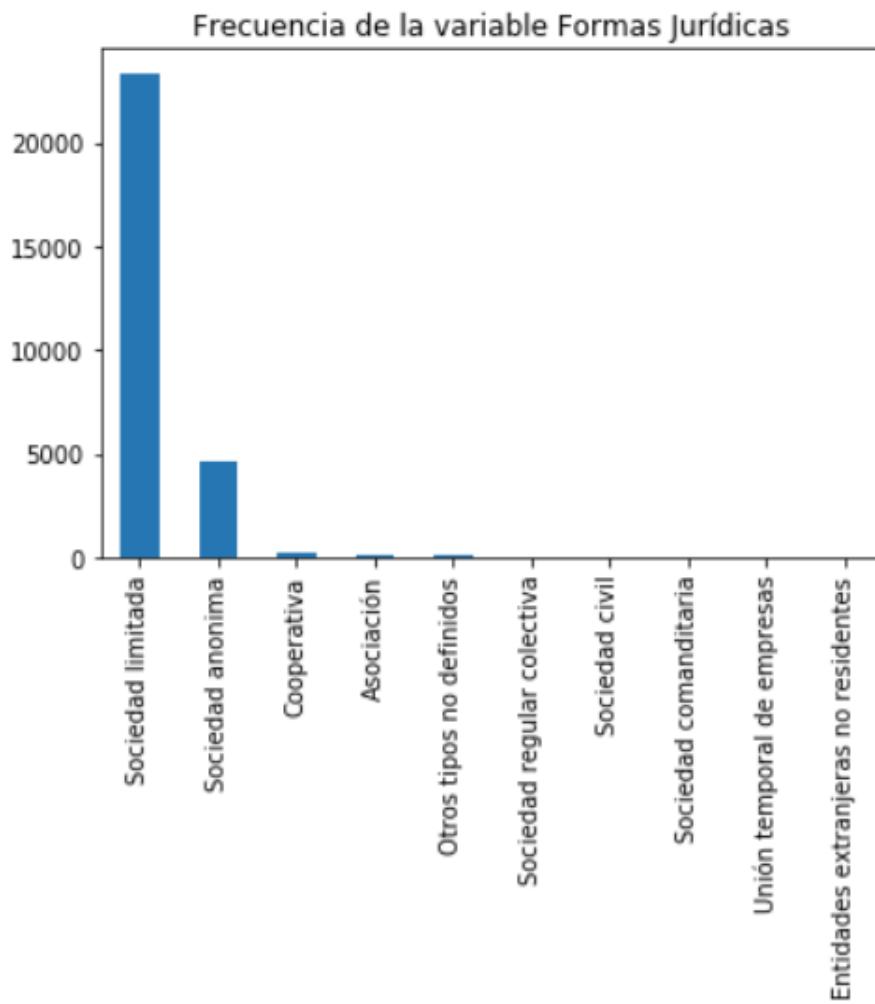


Figura 4.22: Número de empresas según su forma jurídica.

Forma Jurídica	Número de empresas	% Total
Sociedad limitada	23351	82.64
Sociedad anónima	4575	16.19
Cooperativa	222	0.78
Asociación	37	0.13
Otros tipos no definidos	56	0.19
Sociedad regular colectiva	3	0.01
Sociedad civil	2	0.007
Sociedad comandataria	3	0.01
Unión temporal de empresas	3	0.01
Entidades extranjeras no residentes	2	0.007

Tabla 4.6: Frecuencia de las formas jurídicas.

Una vez se definen todas las variables, llega el momento de la selección, donde tenemos 4 formas de elegir las variables:

- Según la literatura: Es decir escoger las 21 variables de la tabla [2.2](#).
- Métodos *filter*: Los métodos *filter*, son métodos en los que a partir de una función de evaluación para un subconjunto S de variables obtenemos una medida numérica para comparar diferentes subconjuntos de variables. Este valor se obtiene a partir de medidas estadísticas como las distancias o medidas de separación entre clases, dependencias entre variables o el estadístico χ^2 .
- Métodos *wrapper*: En este tipo de métodos lo que se hace es evaluar un conjunto de atributos S , a partir de la tasa de acierto obtenida por un algoritmo de aprendizaje. Realmente se trata de una selección de modelos.
- Métodos *filter+wrapper*: Debido a que la selección *wrapper* es compleja en tiempo, lo que se puede hacer es una primera selección a través de métodos *filter*, y una vez reducido el número de variables se utiliza métodos *wrapper* para una última selección.

En cuanto, a la selección de estrategias de búsqueda tenemos los algoritmos secuenciales y los metaheurísticos. En este caso usaremos métodos secuenciales que son aquellos en los que añaden o eliminan atributos al subconjunto de variables de forma secuencial. Utilizaremos este tipo de algoritmos ya que son algoritmos voraces, muy sencillos de implementar y que obtienen muy buenos resultados. Existen dos tipos de selección de estos algoritmos:

- Selección hacia adelante: Se comienza con un conjunto $S = \emptyset$ y de manera secuencial añadimos al subconjunto S el atributo X_i que maximiza la $f(\{S \cup X_i\})$. Es más rápido al evaluar subconjuntos pequeños y tiende a evaluar menos subconjuntos.
- Selección hacia detrás: Se comienza con conjunto completo de las variables $S = \{X_1, \dots, X_n\}$, elimina el atributo $X_i \in S$ que produce el menor decremento en $f(\{S - X_i\})$. Es más lento, al evaluar subconjuntos más grandes, de hecho no se puede utilizar con conjunto de datos grande.

Método Filter

En este caso para el método *filter*, se utilizará como función de evaluación la aproximación de la información mutua de [Battiti \(1994\)](#):

$$I(S; Y) = \sum_{i=1}^n i(X_i : Y) - \beta \sum_{i=1}^{n-1} \sum_{j=i+1}^n I(X_i; X_j)$$

La información mutua entre dos variables mide la dependencia que hay entre ellas o la información que comparten. Puesto que evaluar la información mutua que comparte un subconjunto S respecto de una variable, es muy costoso computacionalmente, una solución es aproximarla a partir de pares.

Esta función, lo que hace es sumar la información mutua que comparte cada variable con la variable clase del subconjunto S , para posteriormente restar la información mutua que comparte cada variable con todas las demás del subconjunto S . El parámetro β sirve para ajustar la importancia que se da a la información mutua que comparten entre si las variables del subconjunto S .

En cada iteración no es necesario, recalcular otra vez el subconjunto entero. Al tratarse de un algoritmo voraz, en cada iteración simplemente se calcula la información mutua que comparte esa nueva variable con la variable clase y se suma de la información mutua que comparte esa variable con las del conjunto de variables S , teniendo en cuenta la beta.

En cada iteración nos quedamos con la variable que maximice la función de Batti, en caso de en una iteración no mejorar la información mutua aportada ya por el subconjunto S no seguimos explorando el resto de variables.

En cuanto a la forma de seleccionar la mejor β para el método *filter*, se hace una comparación de manera secuencial de un conjunto de valores para el parámetro β con la evaluación del modelo Naive Bayes ya que es un modelo muy eficiente, en cuanto a su coste computacional.

Método Wrapper

Los métodos *wrapper* con búsqueda hacia adelante funciona de la misma manera que los métodos *filter* multivariados, simplemente cambia que la función de evaluación es el acierto aportado por ese modelo.

Para el método *wrapper* el modelo para el evaluar el subconjunto de variables sera el modelo de Árbol de Decisión de la librería de *scikit-learn*¹¹, ya que es un modelo bastante potente, y sencillo de extraer reglas en cuanto a su estructura. La parametrización del Árbol de Decisión será el siguiente:

- Mínimo de registros por hoja para la división: 200

¹¹<https://scikit-learn.org/>

- Mínimo de ejemplos para considerarse una nueva hoja: 100
- Profundidad máxima del árbol: 7
- Pesos de las clases: Para la clase concurso 3 y para la variable activas 1. Se ha decidido hacer una clasificación basada en costes puesto que es una solución al tener un problema tan desbalanceado, primando la clase mayoritaria.

Esta parametrización esta hecha de forma que el árbol de decisión sobreajuste a los datos de entrenamiento lo mínimo posible para que los resultados de la selección de las variables en métodos wrapper seán generalizables a otros modelos.

Hay que mencionar, que en cuanto a la selección de variables no se hará sobre todas las variables, si no solo con las de un ejercicio, puesto que queremos guardar la temporalidad de estas una vez seleccionadas.

Estos son los resultados en cuanto a número de variables:

- Método Filter: 66.
- Método Wrapper: 10.
- Método Filter+Wrapper: 10.
- Selección de los expertos: Las 21 más usadas en una recolección de diversos estudios de este campo. Estas variables han sido usadas en diversos estudios como se puede ver en la tabla 2.2.

El método *filter* escogió una β muy baja, lo que hizo que escogiera un tamaño de subconjunto de variables bastante grande. En cuanto al método *wrapper* y *filter+wrapper* seleccionaron las mismas variables, porque el conjunto de las variables *filter* contenía el de las variables que se había usado en el método *wrapper*, con lo cual se hizo una misma selección con diferencia que el método *filter+wrapper* fue más eficiente en cuanto a tiempo que el método *wrapper* como veremos posteriormente.

Comentar que para la selección *wrapper*, se utilizará como métrica de evaluación del subconjunto S una métrica que definiremos posteriormente denominada f_{1score} .

Una vez aplicamos, estos métodos vamos a ver primeramente en tiempo, cual de los 3 métodos automáticos es más eficiente. En la figura 4.23, podemos ver como el método *filter* es el más rápido de los 3, de hecho el proceso más costoso es el de calcular la información mutua para todos los pares de variables es $\frac{n^2}{2} + n$, pero el cálculo del subconjunto una vez se tienen las informaciones mutuas es prácticamente instantáneo.

Por otra parte, el método *wrapper* es el más costoso en cuanto a tiempo, este también depende del modelo con el que evalúes el subconjunto de variables. Por ultimo, tenemos la selección de variables mixta de *filter+wrapper*, donde como es lógico el tiempo usado es el intermedio entre los dos, esto ocurre puesto que primero el método filter lo que hace es reducir la variabilidad de nuestro conjunto de datos de una manera considerable, para posteriormente que el método wrapper en cada iteración no tiene que hacer una cantidad de evaluaciones tan altas como en el método *wrapper*.

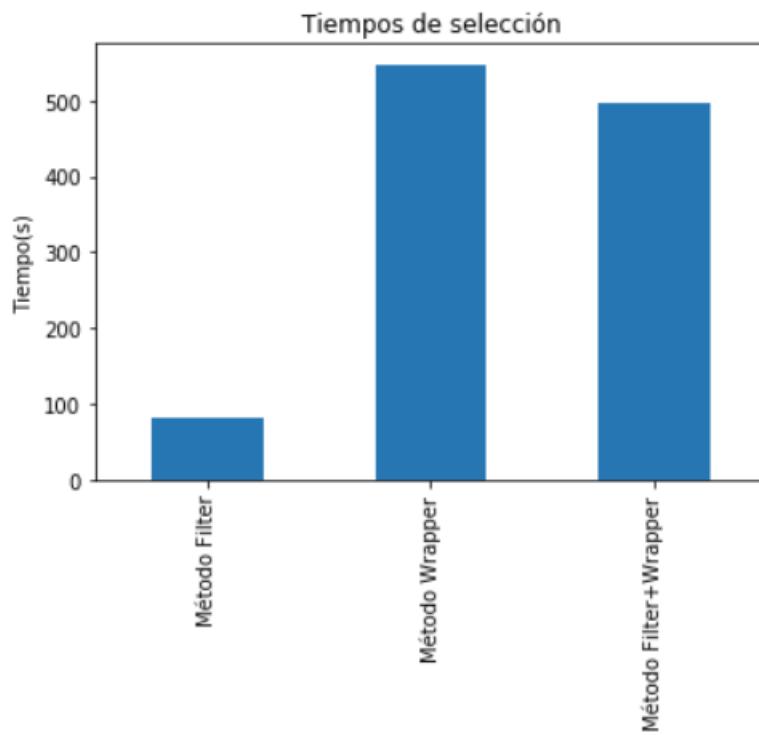


Figura 4.23: Tiempos de algoritmos de selección de variables

Una vez comparado los tiempos, vamos a comparar que resultados arrojan los diferentes subconjuntos de variables, usando modelos muy sencillos ya implementados en la librería de ([Scikit-Learn, 2019](#))¹². Los modelos usados son Naive Bayes, Árboles de Decisión y Regresión Logística que son los modelos que en número han sido lo más usados en la literatura del problema a resolver.

En cuanto, a la parametrización del Árbol de Decisión será el mismo de la parametrización de la selección mediante métodos *wrapper*. Naive Bayes no necesita parametrización alguna. En la regresión logística, usamos la siguiente parametrización:

- Regularización: L2, combate el sobreajuste.

¹²<https://scikit-learn.org/>

- Tolerancia: Tasa de aprendizaje.
- Algoritmo de la optimización del problema: L-BFGS ([Caparrini, 2017](#))¹³ es un algoritmo parecido al de Newton que hace un mejor uso de la memoria.
- Máximo número de iteraciones: 3000, es un número bajo de iteraciones. No se quiere crear modelos finales, si no una evaluación del subconjunto de variables.

En la tabla 4.7 tenemos los resultados teniendo en cuenta el acierto global como métrica. Teniendo en cuenta estos resultados parece que el método *filter* y la selección de los expertos son las que mejor generalizan, para diferentes modelos. Ya que los métodos con la técnicas *wrapper* se sesgán las variables al modelo que se usa para hacer la selección, y por ello en modelos como la regresión logística se obtienen resultados notablemente peores.

En cuanto, al conjunto de datos que no hace selección de variables, el modelo del Árbol de Decisión es capaz de obtener el mejor resultado de todos, porque el mismo hace una selección de las variables para discriminar el espacio de clasificación, en cambio para modelos en los que se usa en la predicción todas las variables los resultados son bastante deficientes, lo que nos advierte que un conjunto de variables estaban causando una gran ruido al problema, empeorando las predicciones de los modelos.

	Filter	Sin selección de variables	Método Wrapper	Método Filter+wrapper	Selección por expertos
Naive Bayes	0.844	0.233	0.866	0.866	0.835
Arbol de decisión	0.919	0.927	0.919	0.919	0.808
Regresión Logística	0.835	0.167	0.671	0.671	0.833

Tabla 4.7: Resultados de la selección de variables usando como métrica la precisión.

Como se puede ver en los resultados de la tabla 4.7, en este problema es necesaria la selección de variables, ya que en Naive Bayes y en la Regresión Logística ha habido una mejora respecto al conjunto de selección sin variables. En el caso de los Árboles de Decisión no hay una mejora puesto que ellos mismos hacen una selección de variables.

Pero como explicamos en capítulos anteriores, el resultado del acierto global en este problema no tienen sentido puesto que nuestro problema esta totalmente desbalanceado respecto de la variable clase y el error más costoso es clasificar a empresas que quebrarán próximamente como empresas sanas.

La matriz de confusión en un problema binario tiene la forma de la tabla 4.8

¹³<http://www.cs.us.es/~fsancho/?e=165>

Predicción Real \ Predicción	Activa	Concurso
Activa	Verdadero positivo(TP)	Falso Positivo(FP)
Concurso	Falso Negativo(FN)	Verdadero Negativo(TN)

Tabla 4.8: Matriz de confusión.

Por ello, se procede a considerar una métrica que tenga en cuenta tanto a aquellas empresas que sean clasificadas como empresas sanas cuando realmente son empresas con un estado fiscal de concurso, como aquellas que son clasificadas como fracasadas cuando son empresas que están en activo. La idea era utilizar una métrica con la forma del F_1 score¹⁴, que tiene esta forma:

$$F_1\text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Donde la $\text{precision} = \frac{TP}{TN+FP}$ y el $\text{recall} = \frac{TP}{TP+FN}$. Esta métrica mide justo lo antagónico de lo que queremos medir nosotros por lo tanto nuestra métrica debería medir la tasa de verdaderos negativos $\text{specifity} = \frac{TN}{TN+FP}$ y la tasa de falsos positivos $\text{negative error} = \frac{FP}{TN+FP}$. Entonces nuestra métrica “ f_1 score” tendría la forma:

$$f_1\text{ score} = 2 \cdot \frac{\text{negative error} \cdot \text{specifity}}{\text{negative error} + \text{specifity}}$$

Habiendo definido nuestra métrica ahora debemos implementarla para poder usarla con la librería de Scikit-Learn, para ello proveen un método denominado *make_scoring*¹⁵, al que se le debe de pasar por parámetro una función al que se le pase los valores reales y los valores predecidos que calcule internamente nuestra métrica.

Si evaluamos otra vez los subconjuntos de variables obtendríamos la tabla 4.9, donde se puede observar que los resultados en general son bastante más pobres que en el caso del acierto del modelo. Si es cierto que la selección obtenida a través de métodos *filter* sigue siendo la que mejor generaliza, siguiendo por detrás los métodos *wrapper* y *wrapper+filter*, pero sin embargo la selección de los expertos ha sido la peor con

¹⁴https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

¹⁵https://scikit-learn.org/stable/modules/generated/sklearn.metrics.make_scoring.html

diferencia cuando con la anterior métrica ha dado unos resultados bastante positivos.

	FSS	Sin selección de variables	Método Wrapper	Método FSSr+wrapper	Selección por expertos
Naive Bayes	0.482	0.277	0.472	0.472	0.112
Arbol de decisión	0.760	0.780	0.758	0.758	0.479
Regresión Logística	0.548	0.285	0.400	0.400	0.083

Tabla 4.9: Resultados de la selección de variables usando como métrica la $f_1 score$

Una vez visto los resultados el conjunto de variables seleccionado para el modelado será el producido por el método *filter*, que ha demostrado producir el mejor subconjunto de variables.

4.3.3. Modelado

En esta fase, visualizaremos un Árbol de Decisión implementado con Scikit-Learn figura 4.24, con la misma parametrización que en el apartado anterior, pero con la diferencia de que la profundidad sería menor para que fuera más fácil de analizar. Este es uno de los modelos que le serían presentados al experto para evaluar si el conocimiento adquirido por el árbol concuerda con el conocimiento del experto.

Cuando presentamos al experto el árbol y empezamos a extraer reglas, observamos que una regla con una separación de espacio del clasificación con 0 fallos es que empresas con una deuda financiera mayor a 346,161 euros se clasifican como activas, cosa que no tiene mucho sentido puesto que la deuda es un agravante y no algo beneficioso en una empresa.

Lo anterior se puede deber a que las empresas de gran tamaño, cuando pasan un cierto límite de ingresos deben pasar su forma jurídica de sociedad limitada a sociedad anónima. Las sociedades anónimas pueden autofinanciarse de dos maneras, una es contraer deuda y otra es sacar acciones de la empresa para que gente externa a ella pueda invertir. Esto segundo es algo que las sociedades limitadas no pueden hacer, lo cual es una gran desventaja frente a las sociedades anónimas.

Por ello, las empresas de mayor tamaño como las sociedades anónimas, pueden contraer más deuda sin que esto en principio pueda ser un inconveniente si la cantidad de deuda no es excesiva.

Además, el experto comenta que usar variables en términos absolutos y no relativos, hace que los datos sean incomprensibles a la hora de extraer conocimiento. También

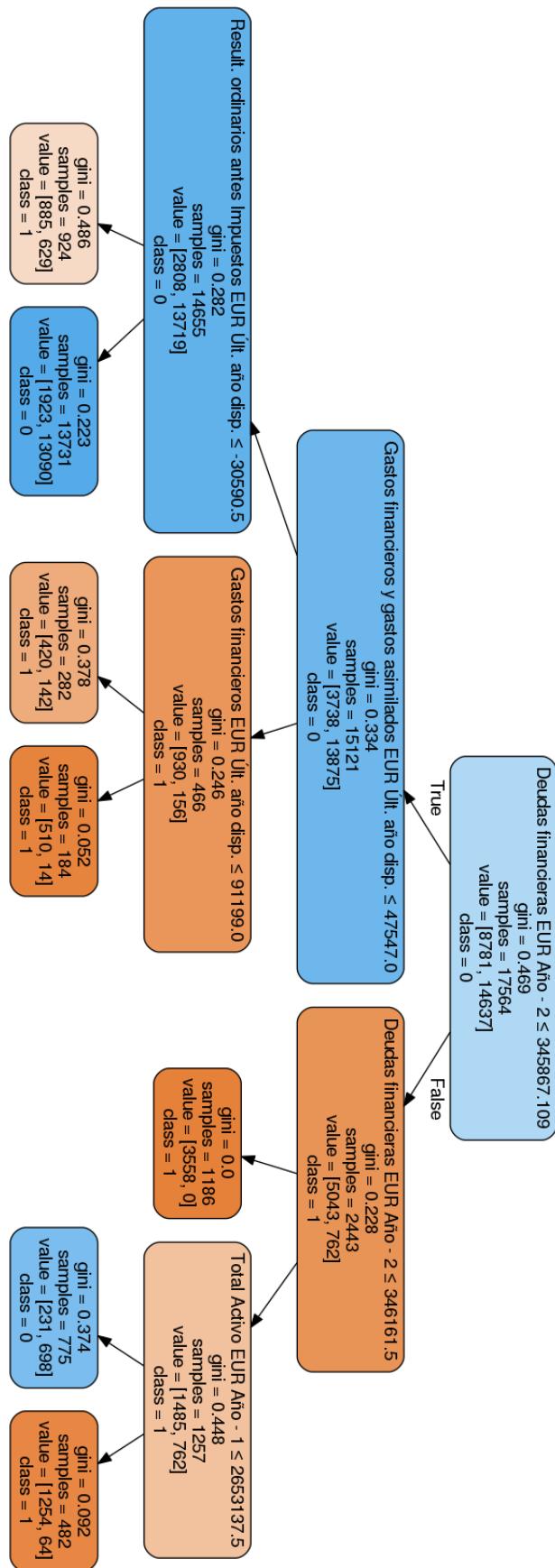


Figura 4.24: Árbol de la cuarta iteración.

comenta que hablar en términos absolutos, para medir si una empresa es sana o, no tiene sentido ya que estamos estudiando empresas con tamaños muy diferentes, donde las cifras de negocio son muy diferentes. Lo único, que estaríamos clasificando es según lo grande o pequeña que sea la empresa en cifra de negocios, si esta se puede considerar como activa o no.

Por ese motivo, en literatura siempre se han utilizado variables relativas como la rentabilidad económica, ya que hablar de fracaso empresarial en términos absolutos en una muestra tan grande no tiene sentido. De ahí que reglas como la anterior surjan como conocimiento extraído por parte del modelo de árboles. Hablar en términos absolutos de los datos financieros de una empresa para medir su riesgo empresarial, se podría hacer en un problema muy acotado donde se estudien empresas de un tamaño y de un sector muy en concreto.

Sin embargo, cuando hemos clasificado con variables que representan con valores relativos, que era la selección de los expertos, hemos obtenido resultados malos con la métrica f_1score , mientras que con el acierto global ha obtenido con todos los modelos alrededor de un 82 %. Si recordamos, la proporción de la variable clase en este conjunto de datos era de 1 empresa en concurso por cada 5 activas aproximadamente, es decir que si todas se clasificaran como 1 obtendríamos un resultado del 80 % de acierto que es un resultado muy parecido al obtenido en nuestros modelos. En cambio, con la métrica f_1score , hemos obtenido una media de alrededor del 22 %, esto puede deberse a que está clasificando casi todo como empresas activas, sin criterio alguno.

Que clasifique de esta manera, no es debido a que los modelos usados sean malos clasificadores. Lo que quiere decir es que la calidad de nuestro conjunto de datos es mala. Esto es debido a que uno de los objetivos de este TFG era ver si no sesgando las muestras y obteniendo muestras de una manera prácticamente aleatoria, usando modelos más potentes que los usados en la mayoría de la literatura se podía generalizar modelos matemáticos sobre el fracaso empresarial, a partir del concepto de fracaso empresarial fiscal.

Por este motivo, la literatura hasta el día de hoy han usado muestras de empresas tan pequeñas. La cantidad de datos disponible sobre empresas que claramente se pueden clasificar como empresas con un estado económico sano, y que entren dentro un tamaño semejante al de otra muestra de empresas con un estado fiscal en concurso, es muy pequeño y difícil de seleccionar.

4.3.4. Puesta en producción

Una vez hemos extraído todas estas conclusiones de esta iteración, para la siguiente y última iteración se hará una selección de la muestra como se ha hecho hasta el día de hoy en la literatura. Además, en cuanto a las variables se usarán las de la tabla 2.2, junto otros coeficientes que vienen ya calculados en el conjunto de variables inicial de SABI.

4.4. Cuarta iteración.

4.4.1. Estudio y comprensión de los datos.

Una vez obtenidas las conclusiones de la iteración anterior, nos fijamos en como hace la selección de la muestra (Alfaro, 2008), donde para las empresas fallidas, escoge aquellas que tengan todos los datos de las variables utilizadas en el estudio. En este caso se escogerán, para las empresas fallidas todas aquellas que tengan los datos de las variables absolutas en el año $(t - 2)$ y $(t - 3)$ antes del fallo empresarial (t). En el caso del año $(t - 4)$ de las empresas fallidas si hubiera valores perdidos se usará el imputador desarrollado en este TFG, ya que teniendo 2 años de los 3 que se usarán en el conjunto de datos el último valor no debería ser difícil de predecir.

Para las empresas activas se hará el mismo proceso respecto de sus años relativos. Comentar que en los dos casos los valores nulos en el año $(t - 2)$ y $(t - 4)$, el número de nulos no es exageradamente alto, pero puesto que el número de instancias había bajado de una manera considerable con respecto las anteriores iteraciones, si seguíamos eliminando instancias por cada variable que presentaría un nulo el conjunto de datos podría haber tenido un tamaño demasiado pequeño.

Además, para las empresas activas nos debemos de asegurar de alguna manera que tienen un estado económico sano, que es algo que no se ha asegurado en anteriores iteraciones. Además, el tamaño de las empresas de la muestra de empresas activas sera semejante a las de la muestra de empresas en concurso, esto último se hará muestreando la clase mayoritaria como en la iteración anterior.

Para asegurarnos que las empresas gozan de un buen estado económico, se exigirá que el BAIT, también conocido como el Beneficio o Resultado antes de impuestos, haya tenido un crecimiento entre el 0 % y el 10 % y que este sea obligatoriamente mayor de 0, todo para los años (t) y $(t - 1)$. Además, el Resultado del Ejercicio debe de cumplir las mismas restricciones que BAIT. Esto no asegura que las empresas gocen de un estado

económico óptimo, pero si puede ser un buen indicativo.

Esto es importante porque el BAIT¹⁶ es el resultado de restar las amortizaciones y el gasto administrativo al Resultado de la Explotación, pero no tiene en cuenta, ni el Resultado Financiero ni los Ingresos Financieros de ese ejercicio. Por este motivo, es importante añadir el Resultado del Ejercicio¹⁷, ya que tiene en cuenta tanto el Resultado Financiero, como el resultado extraordinario¹⁸, que a su vez tienen en cuenta tanto la venta de activos fijos o bonos de la empresa. Esto provocará que haya empresas que posiblemente su BAIT sea positivo pero el resultado de cubrir los pasivos del ejercicio haga que el resultado del ejercicio real sea negativo. Este proceso se describe en la figura 4.25

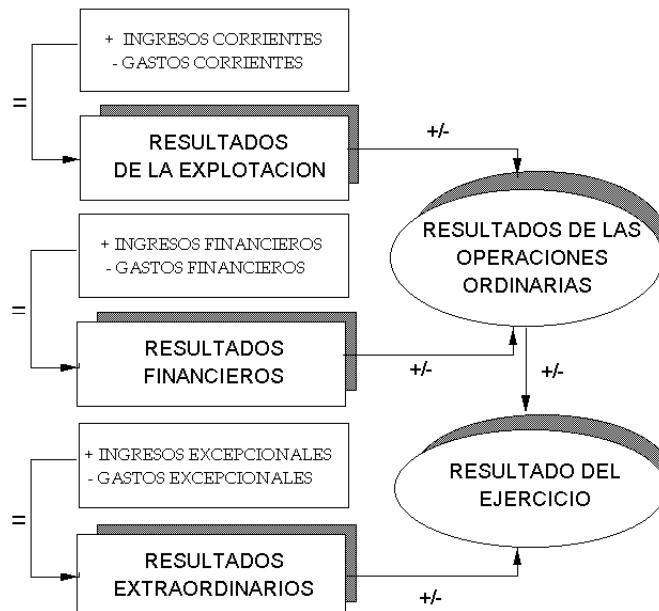


Figura 4.25: Desglose de los resultados.

Después de haber aplicado todas estas restricciones en el caso de las empresas activas como podemos ver en la figura 4.26 que obtenemos un total de 4,337 empresas, que es un número mucho menor que las muestras usadas en las anteriores iteraciones.

En el caso de las empresas en estado de concurso, se aplicaron todas las restricciones de la figura 4.27, obteniéndose un total de 2,416 empresas donde aseguramos que los datos obtenidos por estas sean de calidad, algo que no se aseguró en anteriores iteraciones por no perder instancias ya que la muestra total era demasiado pequeña

¹⁶<https://www.sdelsol.com/glosario/beneficio-antes-de-intereses-e-impuestos-bait/>

¹⁷<http://www.ciberconta.unizar.es/leccion/cf016/200.HTM>

¹⁸<http://www.expansion.com/diccionario-economico/resultados-extraordinarios.html>

ESTRATEGIA DE BÚSQUEDA		<input type="button" value="Guardar"/>	<input type="button" value="Imprimir"/>	<input type="button" value="X"/>	<input type="button" value="Borrar todas las etapas"/>
<input checked="" type="checkbox"/>	1. Estados España: Activa				
<input checked="" type="checkbox"/>	2. Pasivo fijo: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	834.827	834.827		
<input checked="" type="checkbox"/>	3. Pasivo liquido: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	808.193	454.443		
<input checked="" type="checkbox"/>	4. Total Activo: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.464.303	448.882		
<input checked="" type="checkbox"/>	5. Result. ordinarios antes impuestos: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.552.109	448.882		
<input checked="" type="checkbox"/>	6. Gastos financieros: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.452.524	446.794		
<input checked="" type="checkbox"/>	7. Resultado Explotación: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.445.408	445.449		
<input checked="" type="checkbox"/>	8. Fondos propios: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.452.068	445.273		
<input checked="" type="checkbox"/>	9. Activo circulante: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.551.592	445.257		
<input checked="" type="checkbox"/>	10. Cash flow: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.537.364	444.523		
<input checked="" type="checkbox"/>	11. Ingresos de explotación: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.454.723	444.504		
<input checked="" type="checkbox"/>	12. Existencias: Todas las empresas con un valor conocido. Últ. año disponible. Último año -1, para todos los períodos seleccionados	1.253.871	434.773		
<input checked="" type="checkbox"/>	13. Result. ordinarios antes impuestos: its growth rate (%), EUR, entre Último año -2 y Últ. año disponible, min=0, max=10	742.463	255.605		
<input checked="" type="checkbox"/>	14. Resultado del Ejercicio: its growth rate (%), EUR, entre Último año -2 y Últ. año disponible, min=0, max=10	22.842	6.981		
<input checked="" type="checkbox"/>	15. Total Activo (mil EUR): Últ. año disponible, Último año -1, Último año -2, para al menos uno de los períodos seleccionados, min=36.476, max=107.691,199	22.031	4.615		
<input checked="" type="checkbox"/>	16. Result. ordinarios antes impuestos (mil EUR): Últ. año disponible, Último año -1, para todos los períodos seleccionados, min=0	1.423.998	4.585		
<input checked="" type="checkbox"/>	17. Resultado del Ejercicio (mil EUR): Últ. año disponible, Último año -1, para todos los períodos seleccionados, min=0	614.976	4.380		
<input checked="" type="checkbox"/>		617.631	4.377		
		TOTAL : 4.377			
0	Búsqueda booleana	1 Y 2 Y 3 Y 4 Y 5 Y 6 Y 7 Y 8 Y 9 Y 10 Y 11 Y 12 Y 13 Y 14 Y 15 Y 16	<input type="button" value="Actualizar"/>		<input type="button" value="Ver lista de resultados"/>

Figura 4.26: Filtro de empresas activas en la ultima iteración.

frente a las empresas activas.

ESTRATEGIA DE BÚSQUEDA		<input type="button" value="Guardar"/>	<input type="button" value="Imprimir"/>	<input type="button" value="Borrar todas las etapas"/>
<input checked="" type="checkbox"/>	1. Estados Espana; Concurso			
<input checked="" type="checkbox"/>	2. Pasivo fijo: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	5.349	5.349	
<input checked="" type="checkbox"/>	3. Pasivo líquido: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	673.618	3.452	
<input checked="" type="checkbox"/>	4. Total Activo: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.171.802	3.440	
<input checked="" type="checkbox"/>	5. Result. ordinarios antes Impuestos: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.218.522	3.440	
<input checked="" type="checkbox"/>	6. Gastos financieros: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.174.306	3.398	
<input checked="" type="checkbox"/>	7. Resultado Explotación: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.138.375	3.365	
<input checked="" type="checkbox"/>	8. Fondos propios: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.175.324	3.360	
<input checked="" type="checkbox"/>	9. Activo circulante: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.218.333	3.360	
<input checked="" type="checkbox"/>	10. Cash Flow: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.211.790	3.357	
<input checked="" type="checkbox"/>	11. Ingresos de explotación: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.176.156	3.357	
<input checked="" type="checkbox"/>	12. Existencias: Todas las empresas con un valor conocido, Último año -2, Último año -3, para todos los períodos seleccionados	1.053.641	3.112	
		647.672	2.416	
		TOTAL : 2.416		
<input type="button" value="Búsqueda booleana"/>		<input type="button" value="Actualizar"/>	<input type="button" value="Ver lista de resultados"/>	

Figura 4.27: Filtro de empresas en concurso en la ultima iteración.

4.4.2. Estudio de los datos, selección y transformación

Como comentamos en la iteración anterior al no estar estudiando un sector o una muestra de empresas con un cierto tamaño, no es lo más correcto hacer el estudio sobre variables con valores en términos absolutos. Por ello las variables usadas para esta iteración como base serán los ratios seleccionados por todos los expertos y los ratios que incorpora SABI ya calculados en su conjunto de datos.

Para el cálculo de las variables seleccionadas por los expertos, se hace un método que al pasarle el nombre de la columna del numerador y del denominador, crea las 3 columnas correspondientes a los años (t), ($t - 1$) y ($t - 2$) en las empresas activas, y ($t - 2$), ($t - 3$), ($t - 4$) de las empresas en concurso.

De esta manera, solamente se debe concatenar el resultado a nuestro conjunto de datos que esta compuesto por un conjunto de otras 30 variables las cuales están desglosadas según el tipo en la tabla 4.10.

La parte mayoritaria esta compuesta por ratios financieros. Las *dummy variables* que corresponden 3 de ellas a la forma jurídicas que pueden tomar las empresas, y otras 5 que corresponden al tipo de actividad económica desarrollada por la empresa. Periodos, corresponde al tiempo en días de media que tarda la empresa desde el inicio del periodo del cobro del crédito hasta el día del pago total de la cuenta y el tiempo en promedio que tardan en pagar los clientes a la empresa. Por último, el tamaño de la empresa que se mide como el logaritmo natural del Activo Total de las empresas.

Tipo de variable	Items	%Total
Ratio relativos	19	63.3 %
<i>Dummy variable</i>	8	26.6 %
Periodos	2	6.67 %
Tamaño	1	3.33 %

Tabla 4.10: Frecuencia del tipo de variables utilizadas en la última iteración del conjunto de dato SABI.

El conjunto de variables seleccionadas junto con las 21 variables seleccionadas por los expertos y teniendo en cuenta que cada variable de ratios relativos y periodos, corresponde a 3 por cada año relativo y las otras 9, suman un total de 135 variables en el conjunto de datos sin ningún tipo de selección automática.

Ahora, procedemos a hacer diferentes selecciones de variables automáticas de la misma manera que en la iteración anterior. Obteniendo las diferentes cantidades de variables como podemos ver en la tabla 4.11, donde el método *filter* es aquel que ha reducido menos la variabilidad porque la β elegida usando el modelo Naive Bayes ha sido baja como se puede ver en la figura 4.28. El método *wrapper* ha sido aquel que ha disminuido más la variabilidad, al 11.8 % del numero original de variables. Por último, las variables de los expertos como ya conocíamos son 21 variables en total por 3 de los respectivos años relativos usados en los ejercicios, lo que hace un total de 63 variables que representa el 46.6 % del conjunto del problema total.

Método de selección de variables	Número de variables	% Total de variables
Método Filter	74	54.8 %
Método Filter + Wrapper	28	20.7 %
Método Wrapper	16	11.8 %
Selección de los expertos	63	46.6 %
Sin selección	135	100.0 %

Tabla 4.11: Número de variables por método de selección.

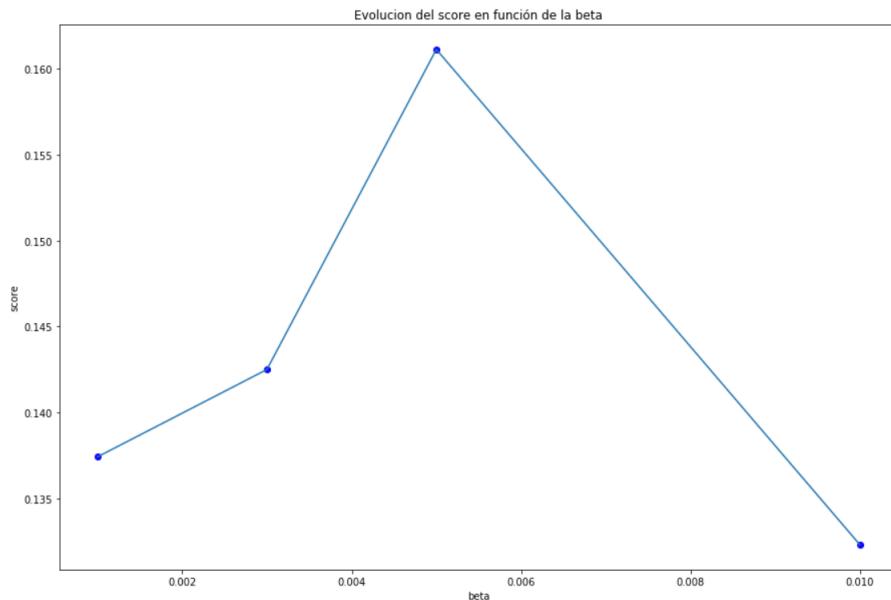


Figura 4.28: Evolución del f_1 score respecto de la β

En la fase de modelado usaremos los mismos algoritmos que en la iteración anterior. Los resultados son los mostrados en la tabla 4.12, que si a priori los comparamos con el acierto en la anterior iteración en la tabla 4.7, pueden parecer que hemos disminuido el rendimiento con los nuevos datos, pero como hemos explicado, el acierto no era la

mejor métrica, puesto que es un problema desbalanceado, aun así los resultados en la regresión logística en media mejora respecto a la iteración pasada.

	Filter	Sin selección de variables	Método Filter+wrapper	Método Wrapper	Selección por expertos
Naive Bayes	0.666	0.664	0.663	0.683	0.668
Árbol de decisión	0.817	0.866	0.824	0.856	0.846
Regresión Logística	0.717	0.725	0.795	0.832	0.800

Tabla 4.12: Resultados de la selección de variables usando como métrica el acierto en la última iteración.

En cambio, si comparamos los resultados de las métrica $f_1 score$ de esta iteración, la tabla 4.13, con los resultados en la tabla 4.9, podemos observar que se mejora de manera notable tanto en los Árboles de Decisión como en la Regresión Logística, teniendo en cuenta que en este conjunto de datos el problema esta mucho menos desbalanceado. Además, lo más destacable es que el rendimiento usando diferentes métricas en el caso tanto de los árboles como de la Regresión Lógistica, se decrementa mucho menos comparado con caso de la iteración anterior.

	Filter	Sin selección de variables	Método Filter+wrapper	Método Wrapper	Selección por expertos
Naive Bayes	0.142	0.131	0.131	0.222	0.148
Árbol de decisión	0.764	0.821	0.773	0.805	0.791
Regresión Logística	0.432	0.492	0.658	0.723	0.688

Tabla 4.13: Resultados de la selección de variables usando como métrica el $f_1 score$ en la última iteración.

Observando la tabla 4.12 y la tabla 4.13, podemos ver como con las métricas del acierto y el $f_1 score$, los mejores conjuntos de variables han sido la selección de los expertos y el método *wrapper*. El método *wrapper*, en concreto ha sido el que mejor rendimiento ha dado para cada modelo en las dos métricas, excepto para los Árboles de Decisión, para el conjunto de datos sin selección de variables, que obtiene mejor resultado en las dos métricas por una diferencia de acierto muy poco significativa en los dos casos. Sin embargo, hay que destacar que la selección de los expertos que era la que daba los peores resultados en la iteración anterior, con mucha diferencia, en esta ha sido la segunda mejor.

Finalmente, el método *wrapper* ha sido capaz de obtener un conjunto de variables que no sesgaba las variables solo a su modelo y ha sido el que mejor resultados ha obtenido para todos los modelos, exceptuando el conjunto de variables sin selección. Esto es así, puesto que los Árboles de Decisión son capaces de hacer su propia selección de variables. Por otra parte, el método *wrapper* ha sido el que más ha reducido

la dimensionalidad del problema por mucho con lo cual será el conjunto de variables usadas para la parte final del modelado, ya que usar conjuntos pequeños de variables en muchas ocasiones ayuda a la generalización del problema.

Las variables seleccionadas en este conjunto final son las adjuntas en la tabla 4.14:

Nombre de la variable	Formula
Tamaño	$\ln(\text{Activo Total})$
Ratio de cobertura de intereses	$\frac{\text{BAIT}}{\text{Gastos Financieros}}$
Rentabilidad sobre el capital empleado	$\frac{\text{BAIT}}{\text{Capital Empleado}}$
Gastos financieros sobre los ingresos de explotación	$\frac{\text{Gastos Financieros}}{\text{Ingresos de explotación}}$
Cash Flow sobre el pasivo líquido	$\frac{\text{Cash Flow}}{\text{Pasivo Circulante}}$
Rentabilidad económica	$\frac{\text{BAIT}}{\text{Activo Total}}$

Tabla 4.14: Conjunto de variables final.

Las variables seleccionadas por el método *wrapper* tienen diferentes lecturas como medida dentro de las empresas:

- Ratio de cobertura de intereses: Este ratio es un medidor del endeudamiento ya que nos informa de la capacidad que tiene la empresa para hacer frente al pago de su deuda, es decir, se está midiendo la capacidad de solvencia de una empresa.
- Rentabilidad sobre el capital generado: También denominado ROCE (Return On Capital Employed), mide la rentabilidad o eficiencia que tiene una empresa para generar beneficios en función del capital que emplea. Este ratio, tiene más sentido usarlo cuando comparamos empresas dentro de un mismo sector, puesto que el capital que se emplea dentro de una empresa de desarrollo software no es el mismo que una fábrica de coches. Aun así, es un buen indicativo de la salud de una empresa puesto que mide cuánto de eficiente es esa empresa para generar nuevos recursos.
- Gastos Financieros sobre los ingresos de explotación: Este tipo de ratio de endeudamiento es una buena combinación con ratios de solvencia, en este caso se indica de qué manera los gastos financieros cubren los ingresos de una empresa, es decir que un valor alto en este tipo de ratios junto con una solvencia mala para afrontar los pagos sería un buen indicativo de una empresa con una salud económica mala.

- *Cash Flow* sobre el pasivo líquido: Este ratio indica que capacidad de caja que tiene una empresa de hacer el pago de sus deudas a corto plazo. Al final empresas que tienen un activo fijo alto pero no es capaz de en un tiempo cercano convertirlo en flujo de caja, no va a ser capaz de hacer frente a sus deudas. Este sería un síntoma de falta de liquidez y por tanto unos de los primeros síntomas del fracaso empresarial.
- Rentabilidad económica: La rentabilidad económica también denominada ROI (*Return on Investment*), mide la capacidad que tienen los activos para generar beneficios, sin tener en cuenta como han sido financiados estos activos. Es diferente del capital empleado que si tiene en cuenta la deuda a corto plazo, ya que es el resultado de la resta del activo total y el pasivo circulante.
- Tamaño: Es un parámetro importante puesto que en muchas ocasiones, la capacidad que tiene una empresa de hacer frente a la deuda es diferente según el tamaño de su activo como hemos comentado anteriormente empresas con activos muy altos y por tanto en muchas ocasiones se les relaciona con empresas con una forma jurídica de sociedad anónima, tienen una capacidad de hacer frente a la deuda a partir de la venta de acciones y de esa manera autofinanciarse. Que es una desventaja frente a la pequeña empresa, ya que la única forma que tienen de autofinanciarse es seguir adquiriendo deuda.

La selección adquirida por los Árboles de Decisión es una selección bastante lógica y que coincide perfectamente con las variables que podría hacer un experto sobre este tipo de problemas ya que tiene en cuenta tanto variables de endeudamiento, rentabilidad económica y estado económico-financiero. Todo esto teniendo en cuenta el tamaño de la empresa.

Para finalizar esta fase abordamos la transformación de estos datos y aunque se pierda esa capacidad explicativa o de extracción de conocimiento vamos a utilizar Algoritmos Genéticos para la extracción de nuevas variables.

Los AGs (Algoritmos Genéticos) son algoritmos de búsqueda, donde el objetivo es a partir de una población inicial de individuos, hacerlos evolucionar a partir de acciones aleatorias como recombinaciones genéticas y mutaciones. De esta manera, podemos encontrar soluciones al problema que no se encuentren cercanas al óptimo local de los individuos iniciales y encontrar otros óptimos locales mejores para nuestra solución del problema.

Para ello, utilizaremos la librería ([GPLearn, 2016](#))¹⁹, que utiliza la programación genética para la transformación de las variables ([Poll, 2005](#)), donde básicamente lo que hace es hacer una búsqueda de nuevas variables generando árboles de operaciones sobre las variables originales como podemos ver en la imagen [4.29](#), donde el intercambio de genes es el intercambio de diferentes subárboles de dos individuos o una mutación podría ser la sustitución de un subárbol por otro de manera aleatoria.

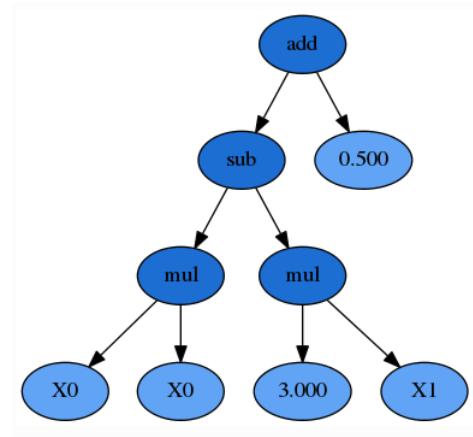


Figura 4.29: Ejemplo de árbol de operaciones generadas por un algoritmo genético.

En el caso del transformador de variables, lo que se utiliza es como función de *fitness* el coeficiente de correlación de Pearson, donde la nueva variable obtenida por el AGs se espera que este directamente correlacionada con la variable clase tanto de manera positiva como negativa.

En el caso del transformador de variables se debe de utilizar el objeto *SymbolicTransformer*²⁰.

Por otra parte los parámetros modificados de los parámetros por defecto serían:

- *generations*: 15, que es el número de generaciones.
- *population_size*: 20,000, el tamaño de la población.
- *n_components*: 12, número de mejores individuos escogidos para formar nuevas variables. Para ajustar este parámetro se hicieron diversas ejecuciones, y se escogió aquel donde la función de evaluación media de la población fuera mayor.
- *max_samples*: 0.9, fracción del conjunto de datos para evaluar un individuo.

¹⁹<https://gplearn.readthedocs.io/en/stable/intro.html>

²⁰<https://gplearn.readthedocs.io/en/stable/reference.html>

- *function_set*: El conjunto operaciones disponibles para la búsqueda serían la suma , resta, multiplicación, división, raíz cuadrada, logaritmo natural, valor absoluto, negada, inversa, máximo y mínimo.

Una vez hecha la transformación de las variables se obtienen individuos como los de figura 4.30, los cuales no se pueden interpretar para extraer conocimiento de los modelos generados por ellos.

```

min(log(Tamano), max(max(sqrt(div(add(inv(Rentabilidad económica (%)) % Año - 2), inv(neg(max(Rentabilidad económica (%)) % Ult. año disp., 0.045)))), sqrt(neg(-0.267))), neg(add(Rentabilidad económica (%)) % Año - 2, Cash flow/Pasivo líquido Año - 1)), max(mul(sqrt(sub(min(Rentabilidad económica (%)) % Ult. año disp., Cash flow/Pasivo líquido Año - 1), neg(Rentabilidad económica (%)) % Año - 2)), sqrt(div(add(inv(Rentabilidad económica (%)) % Ult. año disp., 0.045)))), min(max(-0.039, min(Rentabilidad económica (%)) % Ult. año disp., Rentabilidad económica (%)) % Año - 2), abs(log(Ratio de cobertura de intereses % Ult. año disp.)))), inv(min(neg(max(Rentabilidad económica (%)) % Ult. año disp., 0.045)), Tamano)))) #####
#####
#####
```

Figura 4.30: Ejemplo del individuo de un algoritmo genético.

Ahora vamos a ver cual es su rendimiento respecto a los algoritmos usados hasta ahora para la comparativa de los modelos en la tabla 4.15, donde se puede observar como el rendimiento de todos los modelos sin excepción han mejorado de una manera considerable respecto a la selección de variables *wrapper* que ya era bastante buena.

	Naive Bayes	Árbol de Decisión	Regresión Logística
Acierto	0.9221	0.915	0.926
$f_1 score$	0.889	0.8826	0.892

Tabla 4.15: Resultados del algoritmo genético utilizando el acierto y el $f_1 score$ en una validación cruzada.

En vista de los resultados obtenidos esta transformación de los datos se usará en la fase final de modelado.

4.4.3. Modelado

En la parte del modelado, vamos a empezar con el modelaje de la red neuronal recurrente denominada LSTM, que es un tipo de red neuronal explicada en el capítulo 2 de este documento. Este modelo más complejo, para ello debemos primero elegir la librería con la que la implementaremos. Las posibilidades que tenemos en este caso son:

- ([Tensorflow, 2019](#))²¹: Es una librería de código abierto desarrollada por Google, que ayuda al desarrollo de modelos de aprendizaje automático. En concreto es

²¹<https://www.tensorflow.org/>

famosa por su uso para el desarrollo de redes neuronales y su gran rendimiento para su entrenamiento.

- ([Keras, 2019](https://keras.io/))²²: Es otra librería de código abierto para el desarrollo de redes neuronales, y es muy eficiente porque esta librería está basada en el uso de TensorFlow. Keras tiene una curva de aprendizaje mucho más rápida que TensorFlow, puesto que estaría una capa por encima en el nivel de programación respecto de TensorFlow, aunque no a tan alto nivel como podrían ser librerías como Scikit-Learn.

Por tiempo y sencillez, ya que los resultados que se pueden obtener con Keras pueden ser tan buenos como cuando se usa la librería de TensorFlow se decide usar Keras. Antes de empezar con la implementación del modelo, se debe entender primero, como deben de tratarse de los datos puesto que este tipo de modelo es un tanto especial, en el sentido que lo que estamos usando en este modelo son series temporales y no se trata de modelo de *input-output* clásico.

Lo primero que se debe hacer al usar redes neuronales es normalizar los datos. Para ello utilizaremos el objeto *MinMaxScaler*²³ de Scikit-Learn.

Una vez hemos escalado nuestros datos, debemos convertir los datos en series temporales, esta parte es la más importante y crucial del proceso. Para ello debemos de conocer que es una LSTM, ver ([Olah, 2015](https://colah.github.io/posts/2015-08-Understanding-LSTMs/))²⁴ o ([Karpathy, 2015](https://karpathy.github.io/2015/05/21/rnn-effectiveness/))²⁵ para más detalles. Basicamente, el formato de las LSTM tiene un *input* de la forma [Tamaño del lote, Pasos temporales, Variables], es decir, que estamos trabajando con matrices de 3 dimensiones o de una matriz de matrices. Estos son las 3 dimensiones del *input*:

- Tamaño del lote: Define cuantos ejemplos de entrada queremos que vea nuestra red neuronal antes de actualizar los pesos. Un tamaño de lote pequeño el tiempo de entrenamiento de las redes neuronales puede elevarse de manera considerable de manera considerable. Por otra parte si el tamaño del lote es grande, se reduce la capacidad de generalización de la red neuronal y aumenta el consumo de memoria.
- Pasos temporales: Dice cuantas unidades de series temporales quieres que vea tu red neuronal. Aplicado a este caso, como tenemos 3 ejercicios por parte de cada empresa el numero de series temporales obligatoriamente será 3, puesto que lo que queremos obtener cual será el estado de una empresa habiendo visto los datos financieros de 3 ejercicios diferentes.

²²<https://keras.io/>

²³<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>

²⁴[http://colah.github.io/posts/2015-08-Understanding-LSTMs/](https://colah.github.io/posts/2015-08-Understanding-LSTMs/)

²⁵[http://karpathy.github.io/2015/05/21/rnn-effectiveness/](https://karpathy.github.io/2015/05/21/rnn-effectiveness/)

- Variables: Es el número de variables que se usarán en nuestro problema en total son 16 que corresponden realmente a 5 variables temporales y una fija como es el tamaño, entonces las variables que compondrán la serie temporal serán 6.

Ahora, deberemos convertir nuestro conjunto de datos en una serie temporal, con el formato explicado anteriormente. El proceso es básicamente el diagrama de la figura 4.31, donde pasamos de nuestro conjunto de datos, a tres subconjuntos donde cada conjunto de datos tiene los datos relativos a la secuencia temporal correspondiente, es decir, a los años de los ejercicios (t), ($t - 1$) y ($t - 2$).

Después, como cada subconjunto se encuentra ordenado se añade de manera secuencial una fila de cada subconjunto a otro conjunto de datos nuevo, que lógicamente contendrá un número de instancias igual a número de empresas por 3. Finalmente, se hace con la librería Numpy la función *reshape*²⁶ para cambiar las dimensiones de la matriz final que tendrá de dimensiones [Número de empresas, 3, Número de variables].

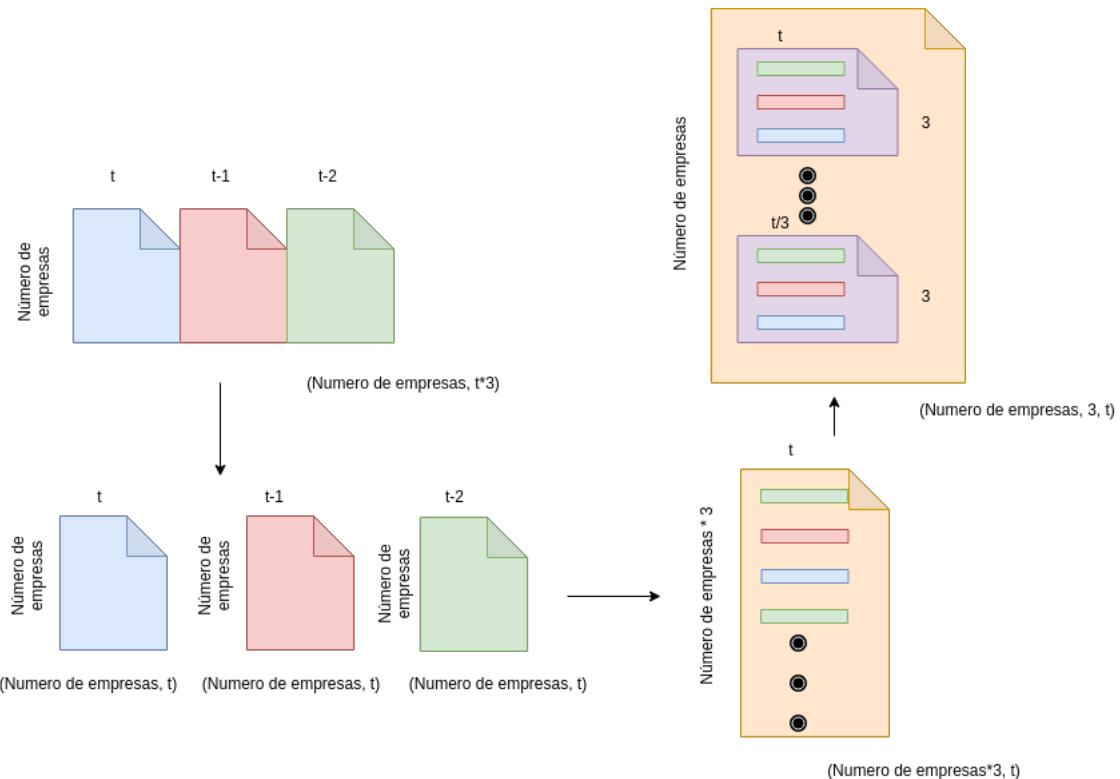


Figura 4.31: Proceso del conjunto de datos a las series temporales.

Una vez tenemos el nuevo conjunto de datos definido, es la hora de implementar la LSTM para ello usaremos el objeto de Keras llamado *Sequential*. Keras también da la posibilidad de aplicarle *Dropout*²⁷, que lo que hace es eliminar de manera aleatoria

²⁶<https://docs.scipy.org/doc/numpy/reference/generated/numpy.reshape.html>

²⁷<https://keras.io/layers/core/#dropout>

conexiones de esa capa. Esto ayuda a combatir el sobreajuste.

Una vez hemos definido todas nuestras capas es necesario definir el algoritmo optimizador de la matriz de pesos en el algoritmo de Propagación Hacia Atrás. Ha habido diferentes mejoras y versiones sobre el descenso del gradiente, una de las más usadas hoy en día es el algoritmo de Adam también llamado *Adaptive Moment Optimization*, donde se calcula el descenso usando el *momentum*, que se utiliza para adaptar la tasa de aprendizaje ([Rivera, 2018](#))²⁸.

En cuanto a estos parámetros se utilizarán los que se aconsejan en la literatura ([Kingma, 2015](#))²⁹, donde se sugiere que el primer momento usado tenga el valor $\beta_1 = 0,9$ y el segundo $\beta_2 = 0,999$.

Además, deberemos definir también las métricas usadas durante el entrenamiento para que época a época nos de información sobre el acierto de nuestro modelo, en este caso se añadirá tanto el acierto, como el acierto con un $threshold = 0,7$. La función de pérdida usada para este problema de clasificación binaria es la entropía cruzada binaria ([Godoy, 2018](#))³⁰³¹.

Comentar que en nuestro modelo para la capa de salida se ha usado la función de activación la función *softmax*, con lo cual la columna de nuestra variable clase deberá de transformarse en un vector de tamaño 2 por cada instancia. Para ello Keras nos provee un método denominado *to_categorical*³², que lo hace de manera automática.

Además, esto se ha implementado para que toda esta implementación no debiera ser cambiada cada vez que se quisiera probar otra estructura diferente de capas ocultas, más o menos como hace Scikit-Learn. Un ejemplo de uso de nuestro modelo sería como en la figura 4.32³³.

En cuanto al rendimiento, no fue el mejor de los modelos. Esto es algo lógico pues el conjunto de datos del que disponíamos era muy pequeño en esta última iteración.

²⁸http://personal.cimat.mx:8181/~mrivera/cursos/optimizacion/descenso_grad_estocastico/descenso_grad_estocastico.html

²⁹<https://arxiv.org/pdf/1412.6980.pdf>

³⁰<https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a>

³¹<https://keras.io/losses/>

³²https://www.tensorflow.org/api_docs/python/tf/keras/utils/to_categorical

³³Todo el código desarrollado en este TFG está en el siguiente enlace: https://github.com/Jaime-TolosaDeLaFuente/TFG_JaimeTolosa

```

lstm_ = RNN(callbacks_list=callbacks_list,metrics=metrics_list,class_weight={0:1.3,1:1})
lstm_.fit(lstm,ylstm,epochs=65,batch_size=30,lr = 0.001,hidden_layers=[8,4],dropout={1:0.3,2:0.3},
validation_data=(val_lstm,val_target))

```

Figura 4.32: Ejemplo de uso LSTM.

Esto se ve claramente en la figura 4.33, donde una vez que la función de pérdida del entrenamiento empieza a converger, la función de pérdida del conjunto de validación comienza a aumentar de manera drástica. Esto ha ocurrido porque lo que está haciendo la red neuronal en ese punto es comenzar a aprenderse los datos de entrenamiento, es decir, se está sobreajustando a los datos.

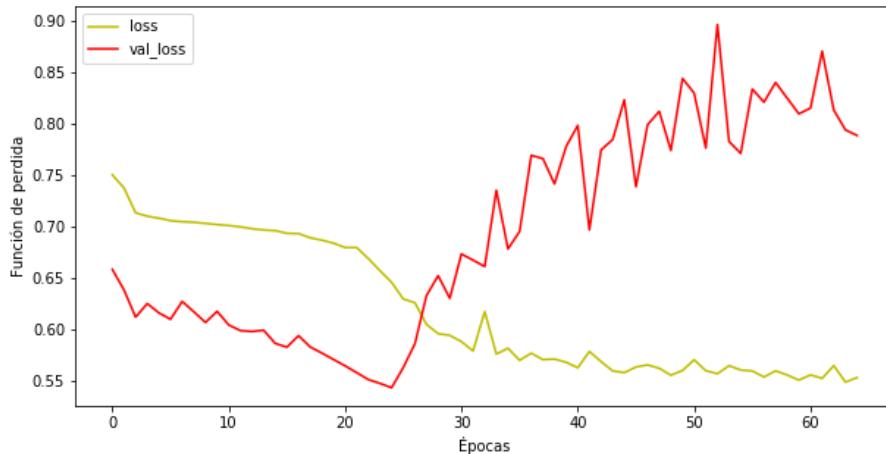


Figura 4.33: Evolución de la función de perdida.

En cuanto al acierto y el acierto con el *threshold*, como es normal se ve el sobreajuste mostrado en la figura 4.33, donde como podemos ver en la figura 4.34 en el momento que la función de pérdida respecto del conjunto de validación comienza a aumentar, el acierto en el conjunto de validación se mantiene o disminuye. Por tanto, el número de épocas que se debería usar para entrenar este modelo es alrededor de las 25. Aun así, el acierto del modelo es demasiado bajo ya que alcanza un máximo en validación de 0.71, que es menor que el acierto aportado por la Regresión Logística y el Árbol de Decisión.

El planteamiento del uso de este modelo aplicado a este problema podría ser bueno, en el caso de poder obtener una cantidad mayor de datos.

En vista, que el resultado de las LSTM no es el esperado, por la pequeña cantidad de datos de la que se dispone, utilizaremos modelos de aprendizaje más potentes que

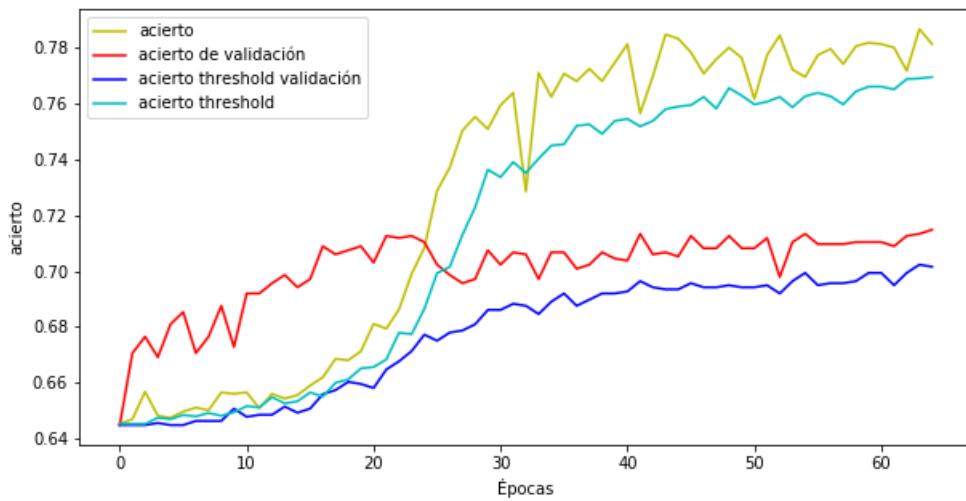


Figura 4.34: Evolución del acierto.

los utilizados en la evaluación para la selección de variables como pueden ser los ensambles basados en árboles que han demostrado ser los modelos con más acierto en las dos métricas usadas en este TFG.

Para ello disponemos de la librería Scikit-Learn que posee todos estos modelos ya implementados y de una manera muy cómoda de usar. Además, usaremos también otro nuevo modelo basados en un Boosting de árboles denominado XGBoost, que ha ganado una gran popularidad en las competiciones de aprendizaje automático y que ha demostrado ser más robusto que otros algoritmos como Adaboost ([Anabel Gómez-Ríos, 2017](#)).

XGBoost implementa el algoritmo de Boosting Gradient. El Boosting Gradient a diferencia de lo que hace el algoritmo AdaBoost que construye árboles de manera iterativa dándole más importancia a aquellas instancias que han sido mal clasificadas. El algoritmo de Boosting Gradient implementa una función de pérdida que intenta minimizar los errores residuales resultantes de la anterior iteración con el error de la siguiente³⁴.

Para ajustar los parámetros de los modelos utilizaremos el objeto *RandomizedSearchCV*³⁵, que sirve para ajustar los parámetros de los modelos de manera automática haciendo un muestreo en el dominio de valores para los distintos parámetros.

³⁴<https://xgboost.readthedocs.io/en/latest/tutorials/model.html>

³⁵https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Para el Árbol de Decisión estos fueron los parámetros optimizados, en el caso del conjunto de datos de las variables seleccionadas por el método *wrapper*:

- *max_depth*:8.
- *min_samples_leaf*:50.
- *min_samples_split*:100.
- *class_weight*:{0:1,1:1}.

Para el Random Forest, estos fueron los parámetros optimizados, en el caso del conjunto de datos de las variables seleccionadas por el metodo wrapper::

- *n_estimator*:100.
- *max_depth*:9.
- *min_samples_leaf*:50.
- *min_samples_split*:100.
- *class_weight*:{0:1,1:1}.
- *bootstrap*:True.

Para AdaBoost³⁶, estos fueron los parámetros optimizados, en el caso del conjunto de datos de las variables seleccionadas por el metodo *wrapper*:

- *estimator*:*DecisionTreeClassifier*(*min_samples_split*= 200, *min_samples_leaf*= 100, *max_depth*= 5).
- *n_estimators*:500
- *learning_rate*:0.01

Para XGBoost³⁷, estos fueron los parámetros optimizados, en el caso del conjunto de datos de las variables seleccionadas por el metodo *wrapper*:

- *n_estimator*:500.
- *max_depth*:8
- *learning_rate*:0.01

³⁶<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>

³⁷https://xgboost.readthedocs.io/en/latest/python/python_api.html

Para la Regresión Logística³⁸, estos fueron los parámetros optimizados, en el caso del conjunto de datos de las variables seleccionadas por el metodo *wrapper*:

- *tol*:0.01
- *solver*:lbfgs.
- *penalty*:l2
- *max_iter*:3000.
- *class_weight*:0:1.5,1:1
- *C*:10

En el caso del Árbol de Decisión, estos fueron los parámetros optimizados, en el caso del conjunto de datos transformados por los AGs:

- *max_depth*:6.
- *min_samples_leaf*:20.
- *min_samples_split*:50.
- *class_weight*:{0:1,1:1}.

Para el Random Forest, estos fueron los parámetros optimizados, en el caso del conjunto de datos transformados por los AGs:

- *n_estimators*:200.
- *max_depth*:7.
- *min_samples_leaf*:600.
- *min_samples_split*:20.
- *class_weight*:{0:1,1:1}.
- *bootstrap*:True.

Para AdaBoost³⁹, estos fueron los parámetros optimizados, en el caso del conjunto de datos transformados por los AGs:

- *estimator*:*DecisionTreeClassifier*(*min_samples_split*= 200, *min_samples_leaf*= 100, *max_depth*= 5).

³⁸https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

³⁹<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>

- *n_estimators*:500
- *learning_rate*:0.01

Para XGBoost⁴⁰ estos fueron los parámetros optimizados, en el caso del conjunto de datos transformados por los AGs:

- *n_estimator*:500.
- *max_depth*:9.
- *learning_rate*:0.01

Para la Regresión Logística⁴¹ estos fueron los parámetros optimizados, en el caso del conjunto de datos transformados por los AGs:

- *tol*:0.001
- *solver*:liblinear.
- *penalty*:l2
- *max_iter*:1000.
- *class_weight*:0:1,1:1
- *C*:20

Ahora usando el conjunto de datos de validación comparamos los rendimientos. En la tabla 4.16, vemos como el Algoritmo Genético obtiene en todos los modelos mejor resultado que la selección de variables mediante el método *Wrapper*. Además no se ven cambios significativos entre modelos, esto puede ser porque la transformación de variables mediante el Algoritmo Genético, lo que busca es que las nuevas variables tengan una correlación lineal respecto de la variable clase por ello modelos como XGBoosting que es capaz de hacer separaciones no lineales del espacio de clasificación, tiene un rendimiento similar o menor a la regresión logística.

Por tanto, una vez aplicamos un Algoritmo Genético para la búsqueda de nuevas variables no son necesarios modelos más potentes que busquen esa separación no lineal puesto que ese trabajo ya se ha hecho en la fase previa de la transformación de las variables.

⁴⁰https://xgboost.readthedocs.io/en/latest/python/python_api.html

⁴¹https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

	Algoritmo Genético		Selección de variables Wrapper	
	Acierto	$f_1 score$	Acierto	$f_1 score$
Árbol de Decisión	0.917	0.877	0.873	0.811
Bosques Aleatorios	0.926	0.890	0.900	0.845
AdaBoosting	0.918	0.878	0.913	0.870
XGBoosting	0.915	0.874	0.913	0.869
Regresión Logística	0.928	0.895	0.817	0.752
Naive Bayes	0.905	0.868	0.677	0.204

Tabla 4.16: Resultados con el conjunto de evaulación

En cuanto a resultados de la tabla 4.16 por parte de la selección de variables *Wrapper*, si encontramos esas diferencias significativas en cuanto a resultado por parte de modelos más potentes respecto de los modelos clásicos como la Regresión Logística, Naive Bayes o los Árboles de decisión. Dentro de los 3 mencionados anteriormente el que mejor ha resultados obtiene de manera destacable son los Árboles de Decisión, aunque su rendimiento es menor en comparativa con los *Ensembles*. En cuanto a estos, los ensembles basados en Boosting son capaces de equiparar su rendimiento a los modelos con el conjunto de datos transformado por los AGs, cosa que el Random Forest no hace, esto se debe a que el error cometido por los modelos no es debido a la varianza si no al sesgo.

Por último, apuntar que como modelo predecir el fracaso una vez entrenados usaremos la Regresión Logística, a partir de un algoritmo Genético, ya que es un modelo sencillo y con un buen rendimiento en coste computacional para la predicción. En cambio para extraer conocimiento de los modelos obtenidos se podría usar el Árbol de Decisión que obtiene unos resultados, más que satisfactorios con el conjunto de validación. Además como veremos usaremos también el algoritmo CN2 como algoritmo de inducción de reglas para compararlo con los árboles y observar si realmente podemos obtener conocimiento de estos dos modelos.

Para finalizar y como conclusión hemos conseguido finalmente utilizando datos para la muestra de empresas en concurso de los años relativos $(t - 2)$, $(t - 3)$ y $(t - 4)$ al fracaso de esa empresa, obtener con una predicción más que satisfactoria predecir anticipar el fracaso empresarial años antes que este se de. Por lo tanto, podemos extraer que los síntomas del fracaso empresarial, pueden aparecer años antes. En el estudio de (Altman, 1968), se obtiene el modelo a partir del año anterior al fracaso de las empresas lo que hace que cuando predice a esas mismas empresas con los datos financieros de

años anteriores al fracaso el modelo pierda mucho rendimiento. Cosa que no ocurre en este caso, por haber tratado el problema desde un enfoque diferente respecto de la definición de fracaso empresarial.

Capítulo 5

Comparación de modelos y el conocimiento del experto

En este capítulo compararemos el conocimiento extraído a partir de los modelos y el conocimiento de los expertos. Por ello, los modelos utilizados en esta sección deben de ser fácilmente entendibles al presentar los resultados. Esta, es una de las partes más importantes del proyecto pues nos ayudará a entender mejor cuales son las causas del fracaso empresarial.

5.1. Árboles de Decisión

En este caso, la profundidad del árbol no debería ser demasiado profunda puesto que lo que se pretende es presentar unas reglas no demasiado complejas a una persona, que en un principio no tiene conocimiento sobre Sistemas Basados en Reglas, por tanto la configuración de parámetros no debería de ser demasiado compleja:

- *min_samples_split*: 50.
- *min_samples_split*:20.
- *max_depth*:4.

La visualización del árbol sería la de la figura 5.1, donde se puede observar patrones de los datos donde en aquellas empresas con un ratio de cobertura de intereses menor del 1 %, pudiendo ser negativa y con una rentabilidad económica menor que 0 %, que también puede ser negativa, se clasifica como una empresa en concurso. Esto es algo lógico pues si la solvencia para hacer frente a la deuda es negativa, es decir que no se puede hacer frente a los pagos financieros y además los activos no son capaces de generar beneficios, ocurre que la empresa no sea capaz de seguir con su actividad económica.

Por otra parte, si la empresa en el último año no es solvente para hacer frente a sus gastos financieros pero sus activos si son capaces obtener rendimiento y en anteriores años la rentabilidad económica de la empresa es positiva, la salud de la empresa es suficiente para mantener su actividad económica. Sin embargo, si en años anteriores los activos de la empresa no han sido capaces de obtener un beneficio suficiente, esa empresa se declara en concurso. Esto anterior ocurre porque si los beneficios no son capaces de hacer frente a los gastos financieros y durante varios años los activos no son capaces de generar beneficios suficientes llegara un punto de no retorno en el que tu empresa no puede seguir acumulando deuda.

El caso anterior para empresas grandes indique que si los gastos financieros son bajos frente a los ingresos financieros, estas empresas pueden entonces hacerse cargo del pasivo. En el caso que los ingresos de explotación superen en proporción a los gastos financieros si se podrá seguir con actividad económica, porque la empresa tiene la capacidad generar los ingresos suficientes para hacer frente al pasivo.

Para empresas con un ratio de cobertura de intereses positivo durante varios años y con una rentabilidad económica también positiva, claramente se declara como una empresa que puede hacer frente a su pasivo, donde sus activos son capaces de generar

beneficios. Este sería un tipo de empresas que en un principio no deberían tener problema alguno para seguir con su actividad económica.

Para empresas que algunos ejercicios su ratio de cobertura ha sido positivo, pero en cambio en otros años sus activos no han sido capaces de generar beneficio, si no que además se ha perdido dinero son empresas que están destinadas al fracaso. Esto ocurre, ya que los gastos financieros no es el total del pasivo en una empresa y no es un indicativo inequívoco de que las empresas gocen de una buena salud, ya que hay gastos como el pago a proveedores que no se ven reflejados en la cuenta de los gastos financieros. Por tanto un motivo como el anterior hace que tus beneficios sean negativos. Para casos como este estaríamos hablando de empresas con problemas económicos.

En el caso que la cobertura de intereses, en el ultimo año haya sido positiva, pero que sin embargo, en años anteriores arrastran un ratio de cobertura muy bajo o negativo, y que además en ese mismo año el rendimiento de los activos ha sido bajo, deja entrever que la capacidad de esa empresa para obtener beneficio es deficiente frente a su envergadura. Con lo cual hablaríamos de empresas que podrían declararse en concurso en un corto plazo de tiempo. Por otra parte, si ese mismo año la rentabilidad económica fue positiva y además el gasto financiero en proporción frente a los ingresos fue bajo, la empresa tiene la capacidad de hacer frente en posteriores años al pasivo.

Sin embargo, si en el último año los gastos financieros han sido mayores en proporción que los ingresos totales aportados por la empresa hablaríamos de una empresa que no es capaz de generar los suficientes ingresos para hacer frente a los pagos.

Todo el conocimiento extraído, a partir de los datos por parte del Árbol de Decisión, coincide con el conocimiento del experto. Por tanto podemos decir que los patrones encontrados por parte del modelo coinciden con la teoría existente sobre el fracaso empresarial.

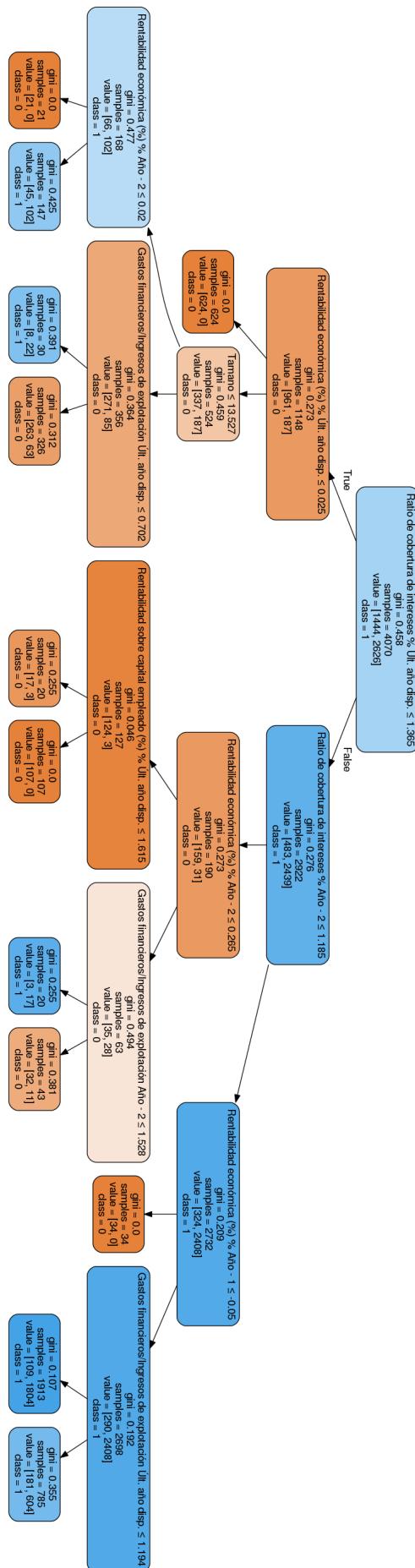


Figura 5.1: Árbol de decisión presentado al experto.

En este caso, hemos tratado con un Árbol de Decisión con una profundidad relativamente pequeña para que su interpretación fuera más sencilla. Pero por la lectura que se obtiene de él, el mayor problema al que se enfrentan las empresas que se declaran en concurso, es que no son capaces de generar el suficiente beneficio para hacer frente al pago de su pasivo.

Además, este problema surge porque el activo de la empresa no es capaz de generar beneficio. A esto, se le suma que empresas que entran en concurso generan un gasto financiero mucho mayor, que los ingresos de explotación de un ejercicio. Por tanto, cuando a esto le aplica los gastos de la explotación el beneficio bruto obtenido es una parte ínfima del gasto financiero, que realmente no es ni el total del pasivo que se debería asumir. Y por ello las empresas llegan a un punto de no retorno donde el fracaso de esa empresa es inevitable.

5.2. CN2. Algoritmo de inducción de reglas.

CN2 es un algoritmo de aprendizaje de inducción de reglas. CN2 esta basado en el algoritmo de AQ(Algorithm Quasi-optimal) y en ID3. El conjunto de reglas es creado por el algoritmo de AQ, pero la forma de tratar los datos como ID3 que es menos sensible a los datos ruidosos.

Para la implementación de este usamos la librería Orange¹ de Python, en la que se debe definir tu conjunto de datos como un objeto tipo Table, donde se tiene que definir a partir de otro objeto denominado Domain cuales son las columnas discretas y las continuas. Ahora, tenemos definido en el formato correcto nuestro conjunto de datos para CN2² de Orange.

La parametrización de este algoritmo fue:

- *beam_width*: 10, el número de flujo de reglas que considera en una iteración, como posible solución.
- *min_covered_examples*: 500, el numero mínimo de ejemplos en una regla. Se escoge, un número así de alto para que el número de reglas no fuera excesivo, y la interpretación fuera más sencilla.

¹<https://orange.biolab.si/>

²<https://blog.biolab.si/tag/cn2/>

Las reglas resultantes del algoritmo CN2 son las mostradas en la figura 5.2, de donde podemos sacar las siguientes conclusiones:

```

Regla 1 IF Rentabilidad económica (%) % Año - 2<=0.01 AND Gastos financieros/Ingresos de explotación Año - 2>=0.0007620847
      590669034 THEN Estado=Concurso [0, 505]
#####
Regla 2 IF Gastos financieros/Ingresos de explotación Últ. año disp.<=0.6693239242692633 AND Cash flow/Pasivo líquido Año
      - 1>=34.40640109710045 AND Cash flow/Pasivo líquido Año - 2>=25.290901269548954 AND Rentabilidad económica (%) % Últ. año disp.>=1.6 THEN Estado=Activa [509, 0]
#####
Regla 3 IF Ratio de cobertura de intereses % Año - 2>=7.2 AND Ratio de cobertura de intereses % Últ. año disp.>=3.0 AND Cash flow/Pasivo líquido Últ. año disp.>=3.641425595073474 AND Ratio de cobertura de intereses % Año - 1>=5.76 AND Rentabilidad sobre capital empleado (%) % Año - 1<=23.41 THEN Estado=Activa [525, 10]
#####
Regla 4 IF Ratio de cobertura de intereses % Últ. año disp.>=2.04 AND Rentabilidad sobre capital empleado (%) % Año - 2<=9.15 AND Rentabilidad sobre capital empleado (%) % Últ. año disp.<=11.13 AND Ratio de cobertura de intereses % Año - 1>=1.2 AND Ratio de cobertura de intereses % Año - 1<=9.6 THEN Estado=Activa [503, 17]
#####
Regla 5 IF Ratio de cobertura de intereses % Últ. año disp.>=1.56 AND Tamaño<=13.868882906287306 AND Rentabilidad sobre capital empleado (%) % Últ. año disp.<=30.32 AND Rentabilidad económica (%) % Año - 1>=0.01 AND Ratio de cobertura de intereses % Año - 2>=1.24 THEN Estado=Activa [474, 56]
#####
Regla 6 IF Ratio de cobertura de intereses % Últ. año disp.<=1.15 AND Cash flow/Pasivo líquido Año - 2>=0.2337004478254715
      6 AND Gastos financieros/Ingresos de explotación Año - 2>=0.1172405930205562 THEN Estado=Concurso [73, 442]
#####
Regla 7 IF Rentabilidad sobre capital empleado (%) % Últ. año disp.>=-1.55 AND Ratio de cobertura de intereses % Año - 1>=9.03 AND Ratio de cobertura de intereses % Año - 2<=30.41 AND Rentabilidad sobre capital empleado (%) % Año - 1<=8.77 AND Ratio de cobertura de intereses % Año - 2<=9.92 THEN Estado=Activa [371, 183]
#####
IF TRUE THEN Estado=Activa [2626, 1444]
#####

```

Figura 5.2: Reglas generadas por CN2.

- Regla 1: Se declara una empresa en concurso, si en años anteriores la rentabilidad económica fue prácticamente 0 o negativa, lo que quiere decir que la cantidad de activos es demasiado alta en comparación al beneficio obtenido. Y por tanto, una empresa con tamaño determinado no es sostenible con unos beneficios tan bajos respecto a los activos o que estos sean inexistentes ya que lo que ha habido son perdidas. Además esto haría, que cualquier empresa que contenga un mínimo de gasto financiero o que adquiera cualquier tipo de deuda, no será capaz afrontar el pago, porque sus beneficios son prácticamente nulos.
- Regla 2: En el caso de que los gastos financieros frente a los ingresos de explotación sea una proporción de prácticamente el 0% y que durante varios años se haya tenido una tesorería lo suficientemente fuerte, para que en cierta medida poder hacer frente a las deudas que haya que pagar a corto plazo y se tenga un mínimo de rendimiento por parte de los activos, esa empresa se puede seguir con su actividad económica.
- Regla 3: No concuerda con el conocimiento experto, que una empresa se clasifique como activa en los casos donde rendimientos tan bajos de ratios de cobertura de intereses junto con una tesorería baja respecto a la deuda a corto plazo y con una rentabilidad sobre el capital empleado menor al 23% una empresa tenga los ratios óptimos como para clasificarse como activa. Esto quiere decir entonces, que el problema principal al que se enfrentan las empresas, no es tanto que la

rentabilidad que obtengan sobre su capital pueda ser bajo, si no que en caso de que los gastos financieros empiezan a ser la mayor parte donde vayan a parar los beneficios. Aunque, esto sea cierto, la rentabilidad de los activos no deja ser importante, ya que una rentabilidad sobre el capital empleado sea muy bajo de una manera constante en el tiempo, puede hacer que el beneficio sea tan bajo que por muy bajo que sean los gastos financieros no se le pueda hacer frente. Por tanto, en el caso de que esta rentabilidad sea baja durante un cierto plazo de tiempo, según los datos, no debería ser un inconveniente a la hora de seguir con la actividad económica si esta no se prologan en el tiempo.

- Regla 4, 5: Semejante a la regla 3, son reglas las cuales no tienen sentido si lo comparamos con el conocimiento del experto, pero indican las mismas conclusiones que la regla 3.
- Regla 6: En el caso de que el beneficio en proporción con los gastos financieros sea menor al 1% pudiendo ser incluso negativo en el ultimo año, hace que los gastos acumulados durante años anteriores sean imposibles de enfrentar, aunque se tenga el suficiente dinero en caja para hacer frente a la deuda a corto plazo.
- Regla 7: En este caso tampoco concuerda con el conocimiento del experto. Tiene un acierto del 66 %, es decir, que acierta casi de manera aleatoria, y por tanto no es una buena regla para extraer conclusiones de ella.

El algoritmo CN2, ha tenido peores resultados a la hora de comparar de extraer conclusiones por parte del experto, ya que se clasificaban empresas como activas cuando la información que aportaban sus ratios no era de una empresa con una buena salud económica.

Por otra parte de las reglas que coinciden con el conocimiento experto podemos concluir que empresas con una rentabilidad económica muy baja son muy sensibles a la adquisición de deuda, puesto que son empresas que no son solventes a la hora de obtener beneficios y no podrán hacerse cargo de la deuda adquirida.

En casos, donde el flujo de caja que es capaz de hacer frente a la deuda a corto plazo, y el gasto financiero frente a los ingresos es muy bajo, en años donde la rentabilidad sobre activos sea baja, la empresa podría seguir con su actividad económica normal, porque la empresa dispone de un dinero en caja para poder hacer frente a sus pasivos, pero esto no se podrá sostener de una manera prolongada en el tiempo puesto que es necesario que la empresa comience a generar otra vez beneficios. Realmente el caso de este tipo de empresas, son aquellas donde sus ingresos son altos pero el beneficio neto

obtenido al restar los gastos de explotación es bajo, y por tanto eso hace que su rentabilidad económica sea baja, una solución sería recortar los gastos que tienen durante el ejercicio, para así aumentar el beneficio obtenido al final del año.

Por contra si los beneficios en el ultimo año no son capaces de hacer frente a los gastos financieros, y los ingresos además son pobres frente a los gastos financieros también, un buen flujo de caja podrá hacerse cargo de una parte del pasivo pero no del total, lo que hará que la empresa comience a acumular pasivos que no sea capaz de afrontar.

Capítulo 6

CONCLUSIONES Y PROPUESTAS

6.1. Conclusiones

El objetivo principal de este proyecto, era poder obtener modelos predictivos del fracaso empresarial, que a fin de cuentas es el objetivo final de la metodología usada. Ahora, vamos a enumerar los objetivos que se proponen en el Capítulo 1 de este proyecto para examinar en qué medida estos han sido alcanzados:

- Un correcto preprocesamiento y transformación de los datos: Para el cumplimiento de este objetivo era importante, conocer la naturaleza de nuestros datos y de qué forma se estructuraban en el conjunto de datos. Por ello fue posible crear un modelo, que predijera los datos nulos de la manera más eficiente posible, que fue clave para la varianza de nuestras variables fuera lo mas alta posible, una vez se imputaran los valores nulos.
- Correcto análisis y visualización de los datos: Fue una parte importante para ver de qué manera se distribuían los valores nulos. Además, de la visualización de las nuevas variables creadas a partir de las existentes en el conjunto de datos, para observar si estas podrían tener una buena aplicación en nuestro problema. Aunque, si es cierto que hacer una visualización total del problema no ha sido posible puesto, que la variabilidad existente en nuestro conjunto de datos era demasiado alta.
- Selección de técnicas de minería de datos más adecuadas al problema y obtención de modelos predictivos mediante su aplicación: En cuanto a los modelos seleccionados fueron los más popularmente utilizados en la literatura, que obtuvieron mejores rendimientos. Además se han considerado nuevas técnicas de aprendizaje

automático las cuales nos ayudaron conseguir mejores modelos, que en estudios anteriores.

- Evaluación de los modelos obtenidos: La selección de la parametrización se hizo mediante técnicas de validación cruzada, y una vez que se conseguía una buena parametrización de los modelos, se usaba el conjunto de validación para ver cuales eran aquellos modelos que mejor generalizaban los datos al observar datos que nunca antes habían visto. Esto nos aseguraba que los modelos obtenidos fueran capaces de generalizar los patrones encontrados para toda la población.
- Extraer información del resultado final: Para ello, se utilizaron modelos de Árboles de Decisión y de inducción de reglas, de los cuales se pudieron extraer conclusiones acertadas de porque ocurría este fracaso empresarial.

Además de los objetivos, gracias a la metodología usada en este proyecto iteración a iteración, se mejoraban partes del proceso de minería de datos que pudieran hacer mejorar nuestro modelo.

De hecho, en este campo de investigación los años seleccionados para estudiar el fracaso empresarial en la muestra de concurso, eran los años relativos anteriores justo al fracaso empresarial. Esto hacía que los modelos usados, en el momento de inferir respecto de año más lejanos al fracaso tuvieran un rendimiento inferior al esperado. En este proyecto se usaron los años $(t - 2)$, $(t - 3)$ y $(t - 4)$, para la parte de entrenamiento y que son años lejanos al fracaso empresarial. Sorprendentemente se fue capaz de predecir con un acierto del 92 % con el conjunto de validación, lo que quiere decir que los síntomas del fracaso empresarial comienzan años antes de que este ocurra.

En todos los casos, este fracaso empresarial no es irreversible pero su desconocimiento puede hacer que en un tiempo próximo llegue a un punto de no retorno. Por este motivo, es importante la inversión en este campo del conocimiento.

Puesto que en este caso solo se han estudiado empresas que pertenecen al territorio español, las características por las cuales ocurre el fracaso solo son extrapolables al conjunto de empresas españolas. Se ha visto que el problema principal, es, que se asumen unos gastos de carácter financiero que posiblemente no sean asumibles por las empresas y, por tanto, el grueso del pasivo de ellas sea el gasto financiero, como se ha visto en las variables seleccionadas.

Además, se ve la importancia de que los activos sean capaces de generar beneficio, puesto que un rendimiento bajo en este aspecto puede hacer que, en un tiempo próximo

no se genere el suficiente dinero como para seguir asumiendo los pagos de la empresa. Por ello, la financiación de las empresas tiene que ser acorde a los ingresos que son capaces de generar, y no asumir costes que no se sabe si vas a ser capaz de hacer frente.

Estos síntomas, serían los primeros que aparecerían del fracaso empresarial según la selección y los modelos usados durante este proyecto. Además, son síntomas que muchos autores de la misma manera han destacado como primeros síntomas del fracaso.

6.2. Trabajo futuro

Como hemos comentado, las empresas estudiadas han sido dentro del territorio español, lo que quiere decir que las características del fracaso empresarial posiblemente no sean extrapolables a cualquier país, como factores más importantes del fracaso. Además solo hemos tenido, en cuenta factores microeconómicos, cuando realmente los factores macroeconómicos de un país tienen una influencia muy grande sobre las empresas.

Por este motivo, en posteriores trabajos sería interesante estudiar factores macroeconómicos con empresas de diferentes países e incluir variables como el PIB, prima de riesgo, inflación, etc. Que tienen una gran influencia sobre la economía de los países y por tanto de las empresas.

Un estudio de este tipo podría hacer que tuviéramos una perspectiva más global de porque se produce el fracaso empresarial o cuales son los primeros síntomas de porque ocurre este. Con el aliciente que el número datos sería mayor y podríamos utilizar técnicas de aprendizaje automático más potentes como pueden ser las LSTM, que en este proyecto no han tenido éxito porque el número de instancias que teníamos en el conjunto de datos final era muy bajo, pero es un modelo que encaja con la naturaleza del problema y podría obtener, unos muy buenos resultados.

Bibliografía

Esteban Alfaro. *Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks*. Science Direct, 2008. [48, 56, 74](#)

María Mar Alonso Almeida. *La transparencia de las empresas en internet para la confianza de los accionistas e inversores: Un análisis empírico*. Universidad Autónoma de Madrid, 2009. [59](#)

Altman. *Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks*. Journal of Banking and Finance, 18(3), pp. 505-529., 1994. [25](#)

Edward I. Altman. *Financial ratios discriminant analysis and the prediction of corporate bankruptcy*. The Journal of Finance, 1968. URL <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1968.tb00843.x>. [2, 8, 9, 10, 38, 54, 93](#)

Julián Luengo & Francisco Herrera Anabel Gómez-Ríos. *A Study on the Noise Label Influence in Boosting Algorithms: AdaBoost, GBM and XGBoost*. Hybrid Artificial Intelligent Systems, 268280., 2017. [89](#)

areapymes. Información sobre ratios financieros. <https://www.areadepymes.com>, 2015. [11](#)

Roberto Battiti. *Using Mutual Information for Selecting Features in Supervised Neuronal Net*. IEEE Transaction On Neural Networks, VOL. 5, NO. 4, July 1994, 1994. [65](#)

William H. Beaver. *Financial Ratios as Predictors of Failure*. Journal of Accounting Research, Vol. 4, 1966. URL <http://www.jstor.org/stable/2490171>. [2, 8, 10, 14](#)

BOE. Ley concursal. <https://www.boe.es/buscar/act.php?id=BOE-A-2003-13813>, 2014. [40, 41](#)

Fernando Sancho Caparrini. Entrenamiento de redes neuronales: mejorando el gradiiente descendiente. <http://www.cs.us.es/~fsancho/?e=165>, 2017. [69](#)

economipedia. Enciclopedia de economía. <https://economipedia.com/>, 2019. 11

S.Fuente Fernández. Análisis discriminante. <http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/DISCRIMINANTE/analisis-discriminante.pdf>, Septiembre 2011. 15

M.J. Múres Quintana & Ana García Gallego. *Factores determinantes del fracaso empresarial en Castilla y León*. Universidad de León, 2004. URL <https://dialnet.unirioja.es/descarga/articulo/1976597.pdf>. 38

Daniel Godoy. Understanding binary cross-entropy log loss: a visual explanation. <https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a>, 2018. 87

Google. Inteligencia google. <https://www.google.com/search?q=concepto+de+inteligencia&oq=concepto+de+inteligencia&aqs=chrome.69i57j0j69i6112j0l2.3144j1j4&sourceid=chrome&ie=UTF-8>, 2019. 19

GPLearn. Gplearn. <https://gplearn.readthedocs.io/en/stable/>, 2016. 83

Thomas Hardy. *IA: Inteligencia Artificial*. Polis, Revista de la Universidad Bolivariana, vol. 1, núm. 2, 2001, p. 0, 2001. 19

Hochreiter and Schmidhuber. *Long Short-Term Memory*. 1997. 27

Juan Duarte Vargas Pedro Ramírez Peradotto Jimmy Reyes Rocabado, Carlos Escobar Flores. *Una aplicación del modelo de regresión logística en la predicción del rendimiento infantil*. Estudios Pedagógicos XXXIII, N° 2: 101-120, 2007, 2007. 15

Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks. <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>, 2015. 85

Keras. Keras. <https://keras.io/>, 2019. 85

Jimmy Lei Ba & Diederik P. Kingma. *Adam: A Method For Stochastic Optimization*. ICLR, 2015. 87

Lennox. *Identifying Failing Companies: A Re-valuation of the Logit, Probit and DA Approaches*. Journal of Economics and Business, 51(4), July, pp.347-364, 1999. 13

Mainstream. Definición de inteligencia. <http://www.intelligence.martinsewell.com/Gottfredson1997.pdf>, 1994. 18

Hamza Mamood. Gradient descent. <https://towardsdatascience.com/gradient-descent-3a7db7520711>, 2019. 25

- José Luis Arquero Montaño. *Procesos de fracaso empresarial en pymes. Identificación y contrastación empírica.* Universidad de Sevilla, 2009. URL https://scholar.google.es/scholar?q=PROCESOS+DE+FRACASO+EMPRESARIAL+EN+PYMES.+IDENTIFICACI%C3%93N+Y+CONTRASTACI%C3%93N+EMP%C3%8DRICA&hl=es&as_sdt=0&as_vis=1&oi=scholart. 8
- Numpy. Librería numpy. <https://www.numpy.org/>, 2019. 62
- Christopher Olah. Understanding lstm. <https://colah.github.io/posts/2015-08-Understanding-LSTM>, 2015. 28, 85
- John R. Koza & Riccardo Poli. *Genetic programming.* Genetic Programming. Search Methodologies, 127164., 2005. 83
- RAE. Inteligencia rae. <https://dle.rae.es/?id=LqtyoaQ|LqusWqH>, 2019. 19
- Mariano Rivera. Descenso de gradiente y variaciones sobre el tema. http://personal.cimat.mx:8181/~mrivera/cursos/optimizacion/descenso_grad_estocastico/descenso_grad_estocastico.html, 2018. 87
- Frank Rosenblatt. *The perceptron: A probabilistic model for information storage and organization in the brain.* Psychological Review, 65(6), 386-408., 1958. 13, 20
- Scikit-Learn. Librería scikit-learn. <https://pandas.pydata.org>, 2019. 53, 68
- María Tascón and Francisco Castaño. *Variables y modelos para la identificación y predicción del fracaso empresarial: Revisión de la investigación empírica reciente.* Universidad de León, 2010. URL <https://www.redalyc.org/html/3597/359733642001/>. 8, 10, 11, 13, 17
- Tensorflow. Tensorflow. <https://www.tensorflow.org/>, 2019. 84
- Mónica Tilves. La cantidad de datos se duplica cada año. <https://www.silicon.es/datos-infografia-2333354>, Abril 2017. 20
- Elvira E. Sanmartín Viviana Janeth. La inteligencia. <http://dspace.ucuenca.edu.ec/bitstream/123456789/2326/1/tps626.pdf>, Septiembre 2010. 19

CONTENIDO DEL CD

En el contenido del CD que acompaña a la memoria podemos encontrar los siguientes recursos:

- Memoria del trabajo en el formato PDF dentro del directorio Memoria.
- El código fuente del trabajo se encuentra dentro de las carpetas que comienzan por el nombre Iteración.

Apéndice A

EJEMPLO DE USO DE LA DE LA BASE DE DATOS SABI

Para usuarios de la UCLM podemos usar de manera gratuita esta base de datos a través de la Red Iris. Para ello, accedemos a las bases de datos de la biblioteca y en la pestaña economía clickamos en la base de datos SABI como se ve en la figura A.1

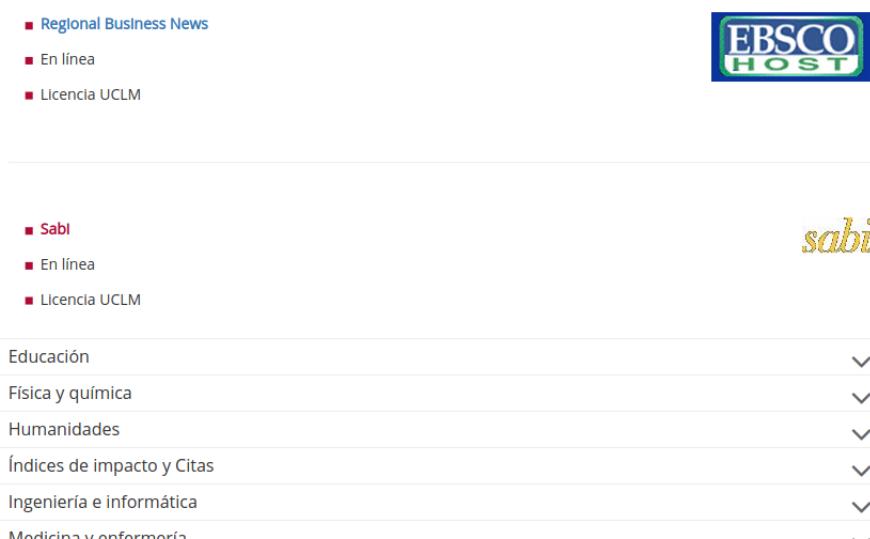


Figura A.1: Bases de datos de la UCLM.

Una vez se accede a la Red Iris debemos elegir la UCLM para acceder a la base de datos de SABI, como aparece en la figura A.2.

Una vez nos identificamos como usuarios de las UCLM nos llevan a la pagina principal de SABI, figura A.3. Donde nos muestran los diferentes filtros que podemos aplicar a la base de datos de SABI.



Figura A.2: Acceso Red Iris.

Figura A.3: Pantalla principal SABI.

Una vez se aplica un conjunto de filtros y clickamos en mostrar resultados. Nos aparecen el conjunto de datos filtrado con las diferentes opciones que podemos aplicar sobre ellos. Para escoger un conjunto de columnas clickaremos sobre la pestaña columnas, como en la figura A.4.

En la parte izquierda tenemos todo el conjunto de variables a elegir, una vez que elegimos una variables no aparece el cuadro de dialogo de la figura A.5

Finalmente, cuando se ha escogido el conjunto de variables final, empezaremos a exportar los conjuntos de datos de 500 en 500 registros. La configuración que yo recomiendo para la exportación es como separador las comas, sin separador de miles, y como extensión el .txt, el resultado quedaría como en la figura A.6

Inicio > Lista (Lista estándar)

Mostrar estrategia de búsqueda

1 de 214 Nota Informe Grupo Columnas Guardar Borrar Alertas Exportar Enviar Imprimir

Las empresas con datos editados se presentan en azul Modificar

Nombre	Código NIF	Localidad	País	Código consolidac.	Último año disponible	Ingresos de explotación mil EUR	Últ. año disp.	Añadir
1. <input checked="" type="checkbox"/> <input type="checkbox"/> CORSAN-CORVIAM CONSTRUCCION SA	A79222709	MADRID	ESPAÑA	U1	31/12/2017	211.236		
2. <input checked="" type="checkbox"/> <input type="checkbox"/> CODYT SA	A58358177	BARCELONA	ESPAÑA	U1	31/12/2009	157.068		
3. <input checked="" type="checkbox"/> <input type="checkbox"/> THYSSEN ROS CASARES SA	A46467965	PUIG	ESPAÑA	U1	30/09/2012	133.968		
4. <input checked="" type="checkbox"/> <input type="checkbox"/> LSB IBERIA SA.	A80447618	CASARRUBIOS DEL MONTE	ESPAÑA	U1	31/12/2014	109.573		
5. <input checked="" type="checkbox"/> <input type="checkbox"/> HOGARES NUEVOS ZARAGOZA SL	B50860485	ZARAGOZA	ESPAÑA	U1	31/12/2006	104.081		
6. <input checked="" type="checkbox"/> <input type="checkbox"/> FICOMSA SERVICIOS FINANCIEROS SL	B84375468	PATERNA	ESPAÑA	U1	31/12/2016	101.949		

Figura A.4: Conjunto de empresas SABI.

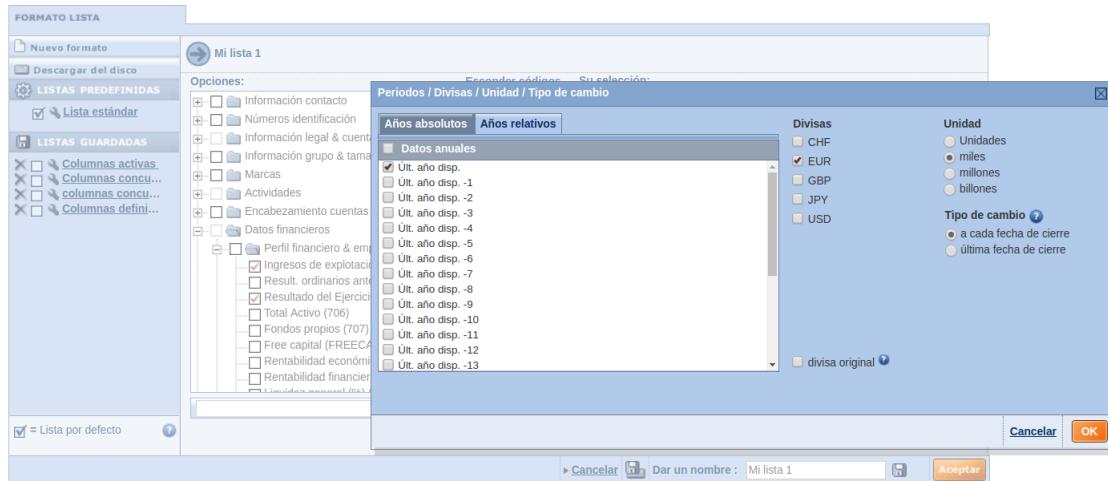


Figura A.5: Unidades de las variables.

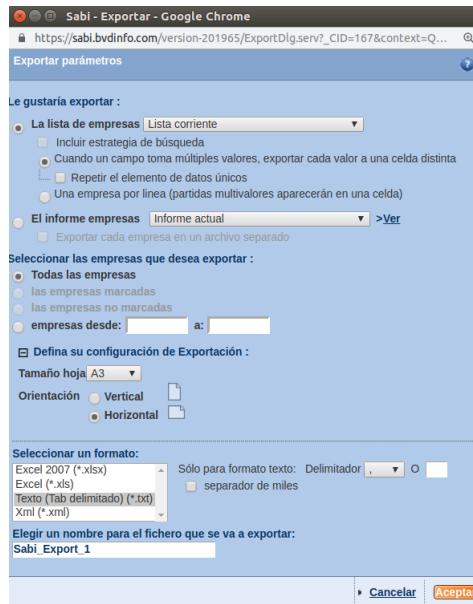


Figura A.6: Cuadro de diálogo de exportación.

Apéndice B

COMPETENCIAS DE LA TECNOLOGÍA CURSADA.

El alumno ha cursado la tecnología específica de computación. En la siguiente tabla se relacionan las competencias de esta tecnología aplicadas durante este Trabajo Fin de Grado .

Competencias	Justificación
Capacidad para tener un conocimiento profundo de los principios fundamentales y modelos de la computación y saberlos aplicar para interpretar, seleccionar, valorar, modelar, y crear nuevos conceptos, teorías, usos y desarrollos tecnológicos relacionados con la informática.	Esta competencia se aplica a la hora utilizar diferentes algoritmos de aprendizaje automático a nuestros datos y generar conocimiento sobre los resultados dados.
Capacidad para evaluar la complejidad computacional de un problema, conocer estrategias algorítmicas que puedan conducir a su resolución y recomendar, desarrollar e implementar aquella que garantice el mejor rendimiento de acuerdo con los requisitos establecidos.	El código desarrollado durante este TFG para la resolución del problema principal ha estado lo mas optimizado posible para que el coste computacional sea el menor.
Capacidad para adquirir, obtener, formalizar y representar el conocimiento humano en una forma computable para la resolución de problemas mediante un sistema informático en cualquier ámbito de aplicación, particularmente los relacionados con aspectos de computación, percepción y actuación en ambientes o entornos inteligentes.	Esta competencia es una de las centrales del TFG puesto que el objetivo ha sido la predicción de un problema que podría ser llevado a cabo por un experto en riesgo financiero.
Capacidad para conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo las dedicadas a extracción automática de información y conocimiento a partir de grandes volúmenes de datos.	Se ha hecho una selección de las técnicas de aprendizaje para poder extraer información de la base de datos SABI y transformarla a conocimiento para ser capaces de clasificar el fracaso próximo de una empresa y obtener conclusiones de porque ocurre esto.

Tabla B.1: Tabla de competencias.