TÍTULO

INSERCIÓN DE MODELOS PREDICTIVOS EN EMPRESAS DEL SECTOR ENERGÉTICO CON MOTIVO DE LA DIGITALIZACIÓN

AUTOR

Luis Guerrero Valor

Esta edición electrónica ha sido realizada en 2023

Tutores	Dr. D. Pedro Javier Zarco Periñán; D. Víctor Jesús Ballesteros Borondo
Instituciones	Universidad Internacional de Andalucía; Universidad de Granada; Universidad de Málaga; Universidad de Almería
Curso	*Máster en Transformación digital de empresas (2021-2022)*
©	Luis Guerrero Valor
©	De esta edición: Universidad Internacional de Andalucía
Fecha documento	2022

TRABAJO FIN DE MÁSTER

Máster en Transformación digital de empresas

Universidad Internacional de Andalucía

Centro: Oficina de Estudios de Posgrado

“Inserción de modelos predictivos en empresas del sector energético con motivo de la digitalización”

Itinerario: Sector Energético

Curso: 2021/2022

Modalidad: Trabajo Técnico

Alumno/a: Luis Guerrero Valor

Director/es:

- Dr. D. Pedro Javier Zarco Periñán
- D. Víctor Jesús Ballesteros Borondo

TRABAJO DE FIN DE MÁSTER

MÁSTER EN TRANSFORMACIÓN DIGITAL DE EMPRESAS

Universidad Internacional de Andalucía

INSERCIÓN DE MODELOS PREDICTIVOS EN EMPRESAS DEL SECTOR ENERGÉTICO CON MOTIVO DE LA DIGITALIZACIÓN

Autor:

Luis Guerrero Valor

Tutores:

Pedro Javier Zarco Periñán

Víctor Jesús Ballesteros Borondo

En colaboración con:

Endesa

Universidad Internacional de Andalucía, 2023

AGRADECIMIENTOS

En este apartado quiero aprovechar para agradecer especialmente el apoyo a todas aquellas personas y entidades que han hecho posible que este proyecto salga adelante.

A la Universidad Internacional de Andalucía, por la organización tan exitosa del Máster en Transformación Digital de Empresas a cuya programación pertenece este proyecto. A su directora y profesores, y en especial a Pedro Zarco, el tutor de este trabajo, por saber ayudarme y guiarme de la mejor manera posible a lo largo del desarrollo del mismo.

A Endesa, por decidir acoger a alumnos como yo, lo cual brinda una oportunidad grandísima de poder formar parte de equipos de trabajo de una compañía de esta envergadura. En especial a Víctor Ballesteros, Raquel Jurado y Manuel Jesús Corchuelo, por ayudarme en mi día a día y hacerme más ameno mi proceso de inserción en el mundo laboral.

A mi familia, mis padres, hermano y pareja, los cuáles me han apoyado una vez más de principio a fin en esta etapa, al igual que en todas a las que me he enfrentado hasta el día de hoy.

SIGLAS Y ACRÓNIMOS

Tabla de siglas y acrónimos.

Abreviatura	Significado Original	Significado Castellano
TFM	Trabajo Fin de Máster	
ANN	Artificial Neural Networks	Redes neuronales artificiales
IA	Inteligencia Artificial	
CUPS	Código Universal de Punto de Suministro	
RAM	Random Access Memory	Memoria de acceso aleatorio
CSV	Coma Separate Values	Valores Separados por Comas
KNN	k-nearest neighbors	k-vecinos más cercanos
SVM	Support Vector Machine	Máquina de soporte vectorial
IoT	Internet of Things	Internet de las cosas
LOOV	Leave one out validation	Dejar uno fuera de la validación
TP	True Positive	Positivo real
FP	False Positive	Positivo falso
TN	True Negative	Negativo real
FN	False Negative	Negativo falso
ATR	Acceso de Terceros a la Red	
NIF	Número de Identificación Fiscal	
CIE	Certificado de Instalación Eléctrica	

LISTA DE FIGURAS

Figura 1.1. Reparto de distribuidoras eléctricas en el territorio español……16
Figura 2.1. Ejes del BIG DATA……28
Figura 2.2. Ciclo de vida de la ciencia de datos. ……..…28
Figura 2.3. Tipos de modelos de Machine Learning en función de su objetivo……30
Figura 2.4. Representaciones de funciones discriminantes……33
Figura 2.5. Esquema de una matriz de confusión……33
Figura 2.6. Esquema de funcionamiento del modelo Random Forest……35
Figura 4.1. Ejemplo de flujo de trabajo en KNIME……40
Figura 4.2. Partes de los nodos en KNIME……41
Figura 4.3. División del interfaz de KNIME……41
Figura 4.5. Flujo implementado para la comparación de modelos……42
Figura 4.6. Metanodo Cross Validation del Random Forest……43
Figura 4.7. Flujo para la parte de aprendizaje del modelo……44
Figura 4.8. Flujo para el lanzamiento del modelo……..44
Figura 5.1. Resultados del modelo del tipo KNN……...49
Figura 5.2. Resultados del modelo del tipo Decision Tree……...50
Figura 5.3. Resultados del modelo del tipo SVM……...50
Figura 5.4. Resultados del modelo del tipo Random Forest……51
Figura 5.5. Resultados del modelo predictivo de este proyecto……..52

LISTA DE TABLAS

Tabla 1.1. Planificación temporal……22
Tabla 2.1. Agrupación de los tipos de modelos de Machine Learning……30
Tabla 2.3. Ideas principales sobre el Machine Learning……...32
Tabla 3.1. Tabla de especificaciones técnicas……..37
Tabla 5.1. Comparación de resultados de modelos predictivos……..47
Tabla 5.2. Fracción de resultados de predicción del modelo……...53
Tabla 5.3. Desglose de costes totales……54
Tabla 5.4. Porcentajes de ahorro en base al umbral seleccionado……55

RESUMEN

El presente documento recoge el desarrollo completo sobre la realización y aplicación de un modelo predictivo, que se encuadra dentro del ejercicio empresarial de las compañías dedicadas al sector energético, en este caso Endesa.

Con el objetivo principal de ayudar a la automatización y gestión eficiente de procesos productivos, se han definido los conceptos teóricos que fundamentan el caso, se ha realizado una organización progresiva de las tareas a llevar a cabo, y se ha terminado por el desarrollo práctico del modelo, teniendo siempre como referente uno ya existente desarrollado e implantado en la compañía. Esto ha permitido comparar los resultados para contrastar así la calidad de los mismos.

En general, se han conseguido alcanzar los objetivos propuestos, logrando poner en marcha un modelo predictivo que ofrece resultados muy positivos, aportando así un punto de valor añadido a la empresa y mejorando la calidad y eficiencia de la actividad de la misma.

Palabras clave: Transformación digital, ciencia de datos, Machine Learning, automatización, energía.

ABSTRACT

This document includes the complete development on the realization and application of a predictive model, which falls within the business exercise of companies dedicated to the energy sector, in this case Endesa.

With the main objective of helping to automate and efficiently manage production processes, the theoretical concepts underlying the case have been defined, a progressive organization of the tasks to be carried out has been carried out, and the practical development of the model, always having an existing one developed and implemented in the company as a reference. This has made it possible to compare the results in order to contrast their quality.

In general, the proposed objectives have been achieved, managing to launch a predictive model that offers very positive results, thus providing a point of added value to the company and improving the quality and efficiency of its activity.

Key words: Digital transformation, data science, Machine Learning, automation, energy.

ÍNDICE

CAPÍTULO 1 - INTRODUCCIÓN. 13
1.1 Motivación del trabajo fin de grado 13
1.2 Objetivos 13
1.3 Contexto 14
1.3.1 Empresa destino. 14
1.3.2 Funcionamiento del sector. 15
1.3.3 Área de contratación. 16
1.3.4 Gestión de rechazos. 17
1.4 Antecedentes. 18
1.4.1 Historia del Machine Learning. 18
1.4.2 Modelos predictivos ya implantados en la compañía. 19
1.5 Resumen de resultados 21
1.6 Planificación temporal 22
1.7 Competencias utilizadas en el TFM 22
1.8 Estructura de la memoria del TFM 24
CAPÍTULO 2 - FUNDAMENTOS TEÓRICOS 27
2.1 Transformación digital. 27
2.2 Ciencia de datos. 27
2.3 Machine Learning. 29
2.4 Aprendizaje supervisado. 32
2.4.1 Clasificación. 32
CAPÍTULO 3 - ESPECIFICACIONES TÉCNICAS 37
CAPÍTULO 4 - METODOLOGÍA EXPERIMENTAL 39
4.1 Obtención de los datos. 39
4.2 Depuración. 39
4.3 Descripción de la herramienta *KNIME*. 40
4.4 Comparación de tipos de modelos. 42
4.5 Implementación del Random Forest. 43
CAPÍTULO 5 - RESULTADOS Y DISCUSIÓN 47
5.1 Resultados de la comparación de modelos. 47
5.1.1 Resumen de resultados, valoración y elección. 47
5.1.2 Resultados del modelo KNN. 48
5.1.3 Resultados del modelo Decision Tree. 49
5.1.4 Resultados del modelo SVM. 50

5.1.5 Resultados del modelo Random Forest. 51
5.2 Resultados del modelo predictivo elegido..... 52
5.3 Presupuesto y análisis financiero. 54
5.4 Tabla de ahorros. 54
CAPÍTULO 6 - CONCLUSIONES Y TRABAJOS FUTUROS 57
6.1 Conclusiones..... 57
6.2 Trabajos futuros. 57
CAPÍTULO 7 - BIBLIOGRAFÍA 59

CAPÍTULO 1 - INTRODUCCIÓN.

1.1 Motivación del trabajo fin de grado

Este trabajo viene motivado por el deseo de conseguir implementar un modelo predictivo que facilite el trabajo diario al equipo de gestión de rechazos. Este equipo es el encargado de operar frente a los inconvenientes que se presentan cuando las operaciones que solicita la comercializadora (como por ejemplo dar de baja o de alta a un cliente) no se pueden realizar por parte de la distribuidora. Esta situación se denomina rechazo, y es el equipo que se comenta el que se encarga de resolverla, pudiendo terminar esta por dos caminos diferentes. Todo esto se explica con detalle a lo largo de la memoria

De este modo, el modelo pretende ayudar en este trabajo, ahorrando aquellas operaciones que son posiblemente automatizables, de modo que los operadores no pierdan tiempo en gestionar esos tipos de rechazos, ahorrando en coste y trabajando de una forma mucho más eficiente.

La motivación principal es conseguir mejorar la experiencia de cliente permitiendo a los operadores centrar sus capacidades en resolver los problemas que verdaderamente aportan valor a los mismos, y no aquellos que son simples y fácilmente solucionables, de los cuales se encargará el modelo.

Todo esto, junto con las ganas de poder llevar a cabo proyectos reales en empresas fuera de lo que propiamente compone una formación académica, hacen que se hayan centrado todos los esfuerzos en realizar un trabajo de calidad y que cumpla con las características que exige el contexto profesional.

1.2 Objetivos

El objetivo del TFM (Trabajo Fin de Máster) es desarrollar un modelo predictivo realizado con la herramienta *KNIME*, impartida en el Máster que presente la probabilidad de que un rechazo de contratación acabe en firme, con el deseo de poder ayudar al equipo coordinador a automatizar el proceso de gestión de rechazos.

Como este se ha ejecutado con datos aportados por la compañía, se podrán comparar los resultados obtenidos con los que ofrece el modelo predictivo actualmente implantado.

En base al objetivo principal, en el anteproyecto se definieron los siguientes subobjetivos:

- Documentar el proceso de negocio donde se va a aplicar el modelo predictivo, identificando los requisitos del modelo y analizando el modelo predictivo actualmente implantado en Endesa.
- Desarrollar del modelo predictivo que cubra los requisitos establecidos mediante la herramienta de programación de alto nivel *KNIME*.
- Identificar y preparar los datos que alimentan al modelo predictivo.
- Ejecutar el modelo predictivo.

- Analizar los resultados obtenidos y comparar con los del modelo predictivo de Endesa realizando un informe de dicha comparativa y las conclusiones que se derivan.

Conforme se ha ido realizando el proyecto, han aparecido tres objetivos nuevos a cumplir que complementan los citados anteriormente:

- Realizar un análisis financiero que constate que el proyecto es viable y abordable.
- Comparar varios tipos de modelos predictivos para justificar que el elegido por la empresa ha sido el más acertado.
- Obtener un producto final que cualquiera puede ejecutar en *KNIME* para predecir rechazos en ocasiones futuras.

1.3 Contexto

En el mundo empresarial existe desde siempre una tendencia a la mejora continua en busca de nuevas formas de desarrollo y nuevos tipos de modelos de negocio con la idea fundamental de hacer crecer la corporación. Por eso, con la llegada de la era digital, las empresas han ido evolucionando para adaptarse a los cambios que esta digitalización exigía.

Es aquí donde surge el concepto de transformación digital, con el objeto de conseguir transformar los procesos para mejorar la eficiencia, aportar más valor a los clientes, disminuir el riesgo y generar nuevas oportunidades de negocio.

Todo ello se fundamenta en los datos, principalmente en la gran cantidad de datos, mejor denominada Big Data, a través de los cuales se consigue la información necesaria para avanzar y desarrollar todo tipo de procesos digitales.

Existe una gran cantidad de variantes y áreas de estudio dentro de la transformación digital, pero una muy importante para las empresas es el área de los modelos predictivos. A través del tratamiento de datos, estos modelos consiguen avecinar los acontecimientos antes de que sucedan, siempre sujetos a una probabilidad de éxito, lo cual permite a las empresas anticiparse para poder afrontar los cambios de una forma eficiente y competitiva. Además, estos modelos permiten el hecho de introducir automatismos en los sistemas cuando se observa que, para una serie de características, el resultado del modelo siempre es el mismo.

Así, este proyecto, vinculado con la formación práctica recibida a través de Endesa, se ha encajado dentro de la parte de la empresa destinada a la comercialización de servicios, más concretamente dentro del área de contratación, en el equipo encargado de la gestión de los rechazos.

Para explicar este encuadre, en los siguientes puntos se desarrolla la información necesaria sobre el sector, la empresa, y en concreto el equipo de gestión de rechazos.

1.3.1 Empresa destino.

El proyecto se encaja dentro del marco de una empresa dedicada al sector de la energía, más concretamente de la electricidad. Esta empresa es Endesa, una de las principales compañías energéticas de España.

Endesa es una empresa líder en el sector eléctrico español y portugués con más de 10 millones de clientes [1], cuyo objetivo principal es contribuir a crear un nuevo modelo energético basado en las energías limpias, el respeto hacia el entorno natural y el desarrollo sostenible. Su prioridad son los clientes, garantizando un servicio de calidad y ajustado a las exigencias de los mismos.

Atendiendo a los últimos años, se ha reforzado la apuesta por la transición energética hacia un modelo descarbonizado, invirtiendo en energías renovables y en la digitalización de la red, manteniendo siempre el foco en mejorar la eficiencia y calidad del servicio a los clientes.

Es por este último punto que este proyecto cobra sentido, en el cual se pretende realzar el proceso de transformación digital en el que está inmersa la compañía a través de la aplicación y puesta en marcha de un modelo predictivo utilizando técnicas Machine Learning.

1.3.2 Funcionamiento del sector.

El sector eléctrico constituye uno de los grandes bloques dentro de los servicios fundamentales que tiene que prestar un país, pues la electricidad es un bien de primera necesidad. Es por ello que este es un sector muy sujeto a normativas y leyes.

El ciclo de vida de la energía empieza por la generación, continúa por la distribución y termina en el consumo de la misma. Entre tanto, las empresas se tienen que encargar de gestionar la energía y ponerla en cada uno de los puntos de suministro, habiendo acordado antes una serie de condiciones de retribución por los servicios prestados a cada uno de los clientes.

El sector de la energía está compuesto por una serie de entidades que, a través de las comunicaciones que establecen entre sí, consiguen que cada uno de los solicitantes tenga energía eléctrica en su hogar o su empresa.

Es por eso que, en general, una empresa de electricidad, en este caso Endesa, se puede dividir en dos grandes bloques: la distribuidora y la comercializadora.

La parte de distribución es la que se encarga como tal de llevar la electricidad hasta todos y cada uno de los puntos de suministro que así lo soliciten, es decir, esta se encarga del diseño, ejecución, puesta en marcha y mantenimiento, de todas y cada una de las infraestructuras, instalaciones y servicios que son necesarios para transportar y entregar la energía eléctrica bajo condiciones de seguridad, eficiencia y calidad.

Por otro lado, la parte de la comercialización es la que se encarga de realizar los contratos a los clientes, sujetos a las distintas tarifas que se ofertan, tanto para particulares como para empresas.

En España existen cinco grandes distribuidoras que cubren la totalidad del territorio nacional que son: Endesa, Iberdrola, Unión Fenosa, EDP (Hidrocantábrico distribución eléctrica) y Viesgo. De este modo, en función de la localización del punto de suministro, este tendrá asignado una distribuidora u otra, como se puede ver en la Figura 1.1 [2].

Figura 1.1. Reparto de distribuidoras eléctricas en el territorio español [2].

Mientras que, por el contrario, existen una gran cantidad de comercializadoras eléctricas, pudiendo el cliente contratar la luz con cualquiera de ellas independientemente de la zona en la que se encuentre. Por supuesto, son cuatro o cinco las empresas que abarcan la mayoría de los clientes españoles, mientras que el resto se los reparten entre las demás. En este caso, Endesa suele ocupar el primer o el segundo puesto en competición con Iberdrola, con unos diez millones y medio de clientes [3].

Es aquí, donde juega un papel fundamental lo que se denomina como "Área de contratación", situada en la parte de la comercializadora y encargada de garantizar la correcta comunicación entre esta y la distribuidora.

1.3.3 Área de contratación.

Como se ha indicado en el punto anterior, esta área es la que se encarga del correcto funcionamiento de las comunicaciones entre comercializadora y distribuidora con el único objetivo de cubrir las distintas operaciones que demandan los clientes.

El recorrido de la información es el siguiente:

- El cliente solicita una operación a través de cualquiera de los distintos canales (presencial, telefónico, web, app, etc.).
- El canal comunica la operación a la comercializadora.
- La comercializadora comunica la operación a la distribuidora.
- La distribuidora ejecuta la operación, siempre que sea posible.
- La distribuidora devuelve la respuesta a la comercializadora.
- La comercializadora activa la operación en sus servidores para poder facturar por los servicios prestados.

Poniendo el foco en las operaciones que se solicitan por parte del cliente, estas se pueden dividir en tres grandes grupos: altas, modificaciones y bajas.

El grupo de altas comprende todas aquellas operaciones que suponen la captación de un nuevo cliente que puede ser, o bien que sea la primera vez que solicita la energía en un punto de suministro, o bien que quiere cambiarse de comercializadora (si se toma el cambio en la dirección de ganar un cliente).

Por otro lado, las modificaciones comprenden todas aquellas operaciones que suponen un cambio dentro de un contrato ya establecido con el cliente en un punto de suministro determinado: solicitud de más o menos potencia, cambio de titular de la cuenta, cambio de tarifa, etc.

Y, por último, las bajas comprenden las operaciones en las que se deja de ofrecer electricidad al cliente en dicho punto de suministro, bien porque ya no necesita electricidad en ese punto, o bien porque quiere cambiarse de comercializadora (tomando ahora el cambio en la dirección de perder al cliente).

Pero hay un punto fundamental en todo este concepto, y es el hecho de que las operaciones no siempre se pueden llevar a cabo. Es por esto que pueden tomar dos caminos: todas las condiciones se cumplen a la perfección y la distribuidora puede llevar a cabo la operación, o se da el caso de que existe algún problema y la distribuidora rechaza la solicitud de la operación a la comercializadora, volviendo a estar en manos de esta última la gestión de dicha operación.

Por eso, el área de contratación tiene fundamentalmente dos tareas que cumplir una vez que recibe las solicitudes de operaciones por parte del canal: garantizar el correcto funcionamiento de la mensajería, y gestionar de la mejor manera posible las solicitudes rechazadas por la distribuidora.

1.3.4 Gestión de rechazos.

Teniendo en cuenta todo lo explicado en los puntos anteriores, es la hora de hablar del equipo de gestión de rechazos, dentro del cual se ha encajado el desarrollo de este proyecto.

Como se ha descrito anteriormente, el rechazo se da cuando se le solicita una operación a la distribuidora y esta no puede ejecutarla.

Los motivos por los que se puede dar un rechazo son extensos y muy diversos, entre los que se pueden destacar:

- Existencia de solicitud previa en curso.
- Necesidad de abrir expediente de acometida.
- Potencia solicitada mayor que la potencia máxima reconocida de extensión.
- Instalación incompleta o con elementos no válidos.
- Acceso imposibilitado más de dos veces por causas ajenas a la distribuidora.
- NIF-CIF no coincide con el del contrato en vigor.
- Falta documentación técnica de baja tensión.

De este modo, la comercializadora puede gestionar el rechazo de dos formas: en firme o reprocesable.

Los rechazos que acaban en firme son aquellos que se refieren a operaciones que no se pueden realizar de ninguna de las maneras por parte de la distribuidora, por lo que se le comunica al cliente que lo que solicita no es posible de cumplir y se termina el proceso; mientras que los reprocesables son aquellos que se refieren a operaciones que vienen denegadas por alguna cuestión que sí tiene solución, poniéndose en contacto con el cliente, o consultando algunos datos en la empresa, etc. De modo que, se hacen las gestiones pertinentes y, una vez arreglado, se vuelve a introducir en el proceso de la mensajería con la distribuidora para que esta vuelva a revisar dicha solicitud.

Al fin y al cabo, todas estas operaciones comentadas requieren una gran cantidad de trabajo a realizar, principalmente porque, al tener tal cantidad de clientes, son cientos de miles las operaciones que se tramitan cada mes. Es por ello que cobra tanto sentido la automatización de los procesos de operación en este tipo de áreas, todo integrado dentro de la digitalización general que está sufriendo el sector.

Así, el desarrollo del modelo predictivo que se describe en este trabajo, pretende facilitar esta tarea de digitalización a través de los resultados que ofrezca sobre la probabilidad de que un rechazo acabe en firme. He aquí, el encuadre del proyecto dentro de esta empresa de tanta envergadura.

1.4 Antecedentes.

1.4.1 Historia del Machine Learning.

El origen del aprendizaje automático en su sentido moderno se suele asociar con el nombre del psicólogo Frank Rosenblatt de la Universidad de Cornell, quien, basándose en ideas sobre el funcionamiento del sistema nervioso humano, creó un grupo que construyó una máquina para reconocer las letras del alfabeto [4], [5], [6].

La máquina, llamada "Perceptrón" por su creador, utilizaba señales analógicas y discretas e incluía un elemento de umbral que convertía las señales analógicas en señales discretas. Esta se convirtió en el prototipo de las modernas ANN (Redes Neuronales Artificiales), y el modelo de su aprendizaje era cercano a los modelos de aprendizaje animal y humano desarrollados en psicología, [7].

A principios de la década de 1960, varios grupos se dedicaron al diseño y prueba de sistemas de reconocimiento de aprendizaje [8], [9], [10].

De esta forma, el marco más sistemático basado en la minimización del riesgo promedio fue desarrollado por Yakov Tsypkin [11], quien demostró brillantemente que este abarca una gran cantidad de algoritmos propuestos por diferentes autores. Aunque la idea de la minimización del riesgo promedio apareció antes en la investigación de operaciones y fue introducida en el control de sistemas inciertos por Feldbaum [12], Tsypkin fue el primero en proponer un enfoque unificado para la adaptación y el aprendizaje. Así, la elección adecuada de la función de coste permitió al autor diseñar diferentes clases de algoritmos descritos anteriormente en la literatura, además de varios nuevos.

En 1969 se publicó el libro de M.Minsky y S.Papert [13] donde se establecían algunas limitaciones por complejidad del problema que podía ser resuelto por los perceptrones. En él, los autores enfatizaron que los perceptrones no podían representar algunas funciones lógicas. Como resultado, se realizó muy poca investigación en esta área hasta aproximadamente la década de 1980. El libro provocó la reducción de la financiación de la investigación de la IA (Inteligencia Artificial) en el mundo durante más de dos décadas. Por eso, este período se denominó más tarde el "Primer invierno de la IA". Sin embargo, aún continuaba el estudio de algoritmos de aprendizaje más complejos.

En 1970-1980 se realizaron más estudios sobre las estructuras y las capacidades de aprendizaje de las redes neuronales multicapa. En 1980, Kunihiko Fukushima propuso una red neuronal convolucional multicapa jerárquica, conocida como "Neocognitrón" [14].

Un impacto significativo tuvo la invención del algoritmo de aprendizaje de retropropagación por parte de varios autores a mediados de los años ochenta [15], aunque sus ideas iniciales se propusieron todavía a principios de la década de 1960, [16].

El comienzo de la primera década del siglo XXI resultó ser un punto de inflexión en la historia de Machine Learning y esto se explica por tres tendencias sincrónicas, que juntas dieron un efecto de crecimiento notable.

La primera tendencia es el Big Data. La cantidad de datos se volvió tan grande que los nuevos enfoques surgieron por necesidad práctica más que por la curiosidad de los científicos.

La segunda tendencia es la reducción del costo de la informática y la memoria en paralelo. Esta tendencia se descubrió en 2004, cuando Google presentó su tecnología MapReduce, seguida de su contraparte de código abierto, y juntos hicieron posible distribuir el procesamiento de grandes cantidades de datos entre procesadores simples. Al mismo tiempo, el costo de la RAM (Random Access Memory) disminuyó significativamente, lo que abrió la posibilidad de trabajar con grandes cantidades de datos en la memoria y, como resultado, surgieron numerosos tipos nuevos de bases de datos, incluido el NoSQL.

La tercera tendencia es el desarrollo de nuevos algoritmos de aprendizaje automático profundo. Después de muchos años de estudio de las redes neuronales multicapa, nació un nuevo concepto, la tecnología de redes neuronales profundas.

Ante todo esto, existe mucha bibliografía que habla sobre la evolución del Machine Learning, siendo importante destacar el elevado ritmo al que avanza el desarrollo de nuevas tecnologías, lo que hace que rápidamente la gran mayoría de artículos de épocas anteriores queden obsoletos y los investigadores tengan que renovarse continuamente para seguir el ritmo de la transformación digital que se sufre en la actualidad.

1.4.2 Modelos predictivos ya implantados en la compañía.

En esta compañía ya existe un modelo predictivo puesto en marcha cuyo funcionamiento se describe a continuación.

Este modelo se implantó con el objetivo principal de que, una vez este haya devuelto las probabilidades de que los rechazos acaben en firme, a decisión del encargado, cargar en el sistema los que se encuentren por encima de cierto umbral para que este los resuelva en firme automáticamente y así eliminar el punto en el que los operadores han de hacerlo de forma manual, para reducir en costes y aumentar en eficiencia.

Pero para ello, el flujo de trabajo que se ha establecido es el siguiente:

1. Extracción de datos de entrada.

El equipo de rechazos extrae de su sistema operativo los rechazos que se encuentran en estado “Pendiente de gestión” en ese momento.

2. Carga de datos en el modelo.

A través de una dirección web se accede al servidor que contiene el modelo y se le cargan los rechazos pendientes mediante un archivo CSV (Coma Separate Values).

3. ¿Está el proceso ya automatizado?

Una vez se cargan los rechazos estos han de superar el primer filtro que es el de si el tipo de rechazo que los caracteriza consta ya como un proceso automatizado o no. Esto sucede porque existen algunos motivos de rechazo que ya se han automatizado para que se resuelvan de una u otra forma debido a que siempre acababan de la misma manera, por lo que el sistema tiene que comprobar si existe algún rechazo que esté dentro ese grupo y eliminarlo de la predicción, pues ya se va a resolver de forma automática.

4. Enriquecimiento de información.

Para que el modelo ofrezca un resultado preciso no le es suficiente con la información de entrada que le proporciona el equipo de rechazos, si no que ha de enriquecerse con información extraída de un servidor online que contiene datos relacionados con los rechazos. Los datos que se solicitan son del estilo del producto que tiene contratado el cliente en cuestión, la documentación que este ha presentado para realizar la solicitud que se ha rechazado, si tiene contratos previos o no, la localización del cups (Código Universal de Punto de Suministro), etc.

5. Predicción.

Este es el momento en el que entra en juego el propio modelo predictivo y devuelve los resultados.

6. Establecimiento de umbrales.

Una vez conocidos los umbrales de probabilidad de que los rechazos sean en firma, el encargado decide cuál será el umbral a partir del cual se van a gestionar de forma automática. Por ejemplo, si se decide que el umbral diferenciador es el 80%, todos los rechazos que tengan una probabilidad de ser en firme igual o mayor al 80% se finalizarán en firme automáticamente, mientras que los que estén por debajo de dicho valor han de ser revisados por los operadores que decidirán el camino por el que finalizarán estos rechazos.

Estando definido el flujo de trabajo, es la hora de hablar del tipo de modelo predictivo utilizado.

El modelo utilizado por el equipo ha sido un Random Forest, cuyas características se explican en el punto 2.4.1.

Es cierto que esta no es la tipología de modelo más compleja que existe, pero el equipo se dio cuenta de que predecía con unos índices de acierto muy elevados. Esta es una decisión que siempre se suele tomar a la hora de elegir cuál es el tipo de modelo que se va a implementar, lo normal es elegir el modelo más sencillo que se pueda siempre que cumpla con una buena calidad predicción.

Con respecto al funcionamiento propiamente dicho, este Random Forest lanza un modelo predictivo para cada tipo de rechazo, haciendo así que sea mucho más efectivo el proceso. En el caso de que un tipo de rechazo no tenga suficientes instancias como para lanzar una predicción, estas se van agrupando en un cajón general y se predecirán juntas posteriormente.

Para optimizar la configuración de los parámetros se han usado las siguientes consideraciones: número de estimadores, mínima división de muestras, máxima profundidad, y mínima cantidad de hojas de muestras.

En general, se constituyó un modelo efectivo y con una gran capacidad de cómputo para devolver valores muy precisos en un periodo corto de tiempo.

1.5 Resumen de resultados

En cuanto a los resultados, existen dos puntos principales a tratar: el resultado sobre la comparación de los distintos tipos de modelos predictivos posibles a utilizar, y el resultado del propio modelo predictivo utilizado.

Con respecto a la comparación, los resultados han demostrado que, frente al modelo del tipo KNN (k-nearest neighbors), el del tipo Decision Tree, y el del tipo SVM (Support vector machine), el Random Forest es el que ofrece el mayor índice de calidad de predicción. Tanto el KNN como el SVM devuelven resultados muy negativos. Es cierto que el Decision Tree ofrece una calidad de predicción muy elevada, pero el Random Forest es casi perfecto, ya que conseguir un modelo que acierte siempre es casi imposible.

Esto justifica la decisión de haber escogido para este proyecto un modelo del tipo Random Forest, que es el mismo que el escogido por la empresa en el lanzamiento de su modelo predictivo.

Por otro lado, en cuanto a la predicción del modelo propiamente constituido a lo largo de este proyecto, la calidad de la misma es muy elevada. Tanto es así que, de 983 rechazos que se habían resuelto en firme, el modelo ha acertado en 972, fallando solo en la predicción de 11 rechazos. Del mismo modo, para los 1884 que se habían resuelto como reprocesables, el modelo acierta en 1854, fallando tan solo en 30.

Está claro que se han conseguido unos resultados muy exitosos, habiendo cumplido los objetivos y llegando a la confección de un modelo sólido y fiable para la predicción de rechazos por parte de la empresa.

1.6 Planificación temporal

En la Tabla 1.1 Se indica el tiempo que se ha tardado en la realización de las distintas actividades que engloban el desarrollo completo del trabajo fin de máster:

Tabla 1.1. Planificación temporal.

Mes	Tareas realizadas	Duración en días
Abril	> Recolección y depuración de datos (1 día) > Repaso de conocimientos sobre el manejo de la herramiena *KNIME (1,5 días)*	2,5
Mayo	> Recolección y depuración de datos (1 día) > Análisis del modelo predictivo ya implantado en la empresa (3 días) > Identificación de configuraciones de la herramienta *KNIME* para el desarrollo del modelo (1 día) > Inicio de redacción de la memoria (2 días) > Desarrollo completo del modelo predictivo (8 días)	15
Junio	> Escritura del resto de la memoria (13 días) > Ajustes del modelo para su correcto funcionamiento (2 días)	15
Julio	> Corrección y modificaciones sobre la redacción de la memoria (1,5 días) > Preparación de la presentación para la defensa (3 días) > Preparación de la defensa (3 días)	7,5

Se han requerido un total de 38 días, trabajando una media de 8 horas diarias, lo que suponen un total de 304 horas invertidas.

1.7 Competencias utilizadas en el TFM

Este tipo de trabajos técnicos, exige a los estudiantes la capacidad para abordar proyectos de dirección, creación y gestión de procesos que ayuden a la transformación digital de las empresas, debiendo estos de ser capaces de afrontar con determinación todas las fases que los componen, así como solucionar de forma eficiente y efectiva cualquier tipo de inconveniente que surja en el proceso.

Por ello, durante la realización de este proyecto, se han desarrollado una serie de competencias que se enumeran a continuación.

Competencias básicas y generales:

- CG1. Interpretar y reproducir el método científico para analizar y formular juicios, bien sean experimentales y/o teóricos, en el ámbito de la Transformación Digital de Empresas.
- CG2. Demostrar dominio en la utilización de bibliografía científica y bases de datos, así como en el análisis de documentos científicotécnicos, en el ámbito de la Transformación Digital de Empresas.
- CG3. Contrastar, revisar y desarrollar informes, presentaciones y/o publicaciones científicas en el ámbito de la Transformación Digital de Empresas.
- CG4. Saber interpretar el marco normativo básico regulador del ámbito de la Transformación Digital de Empresas.

- CG5. Diferenciar y aplicar de forma eficiente las Tecnologías de la Información y la Comunicación en el ámbito de la Transformación Digital de Empresas.
- CG6. Desarrollar un proyecto innovador en el ámbito de la Transformación Digital de Empresas, con iniciativa y una actitud proactiva y ética, asumiendo responsabilidades propias del ámbito profesional, en un entorno multilingüe y multidisciplinar.
- CB6. Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en el contexto de investigación.
- CB7. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
- CB8. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios.
- CB9. Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.
- CB10. Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

Competencias transversales:

- CT1. Mostrar compromiso con el respeto y promoción de los Derechos Humanos, la cultura de la paz y la conciencia democrática, los mecanismos básicos para la participación ciudadana y una actitud proactiva para la sostenibilidad ambiental y el consumo responsable.
- CT2. Examinar los Objetivos de Desarrollo Sostenible, especialmente los relacionados con la promoción del Estado de Derecho en los planos nacional e internacional; la garantía de acceso público a la información y proteger las libertades fundamentales, de conformidad con las leyes nacionales y los acuerdos internacionales; el fortalecimiento de las instituciones nacionales pertinente mediante la cooperación internacional, y la promoción de leyes y políticas no discriminatorias en favor del desarrollo sostenible.
- CT3. Aplicar la igualdad de género y la reducción de desigualdades en la sociedad a través del conocimiento y la educación y desarrollar un compromiso ético como ciudadano y como profesional.
- CT4. Interpretar la información y aplicar el conocimiento de forma crítica.
- CT5. Desarrollar las aptitudes para el trabajo, la comunicación efectiva, la planificación y gestión del tiempo, el esfuerzo, el aprendizaje permanente, la búsqueda de la calidad, así como el espíritu creativo y emprendedor, además del liderazgo, para el adecuado desarrollo de proyectos académicos y profesionales.
- CT6 - Desarrollar las aptitudes para el trabajo, el esfuerzo, la búsqueda de la calidad, así como el espíritu creativo y emprendedor, además del liderazgo, para el adecuado desarrollo de proyectos académicos y profesionales.

Competencias específicas:

- CE1. Diferenciar los procesos empresariales y aplicar las tecnologías, plataformas y herramientas adecuadas para la transformación digital.
- CE2. Aplicar adecuadamente las metodologías de desarrollo e innovación empresarial.
- CE3. Construir visualizaciones de datos que ayuden a la toma de decisiones.
- CE4. Identificar las principales amenazas en los diferentes campos de aplicación y evaluar y gestionar los riesgos asociados.
- CE5. Comparar los servicios, los mecanismos y las herramientas de seguridad y privacidad existentes, y saber aplicarlos, implementarlos o integrarlos en los diversos entornos o escenarios de aplicación, ya sean convencionales o críticos, y de acuerdo con las actuales normativas, estándares y tecnologías.
- CE6. Aplicar métodos matemáticos, estadísticos y de inteligencia artificial para modelar, diseñar y desarrollar aplicaciones, servicios, sistemas inteligentes y sistemas basados en el conocimiento.
- CE7. Analizar datos y extraer información relevante de los mismos.
- CE8. Revisar tecnologías para la implementación de sistemas de gestión y explotación de datos.
- CE9. Diferenciar y adaptar las herramientas, protocolos y plataformas de desarrollo de IoT (Internet of Things).
- CE10. Diseñar, configurar, implementar y evaluar soluciones de computación en la nube.
- CE11. Integrar las tecnologías relacionadas con la informática industrial y las comunicaciones para la mejora de los procesos de producción.
- CE12. Diseñar proyectos de automatización y robotización en el ámbito industrial.
- CE13. Demostrar el conocimiento de las técnicas de fabricación integrada por computador para el desarrollo de un nuevo producto comercial.
- CE14. Aplicar la interacción hombre-robot y robot-robot en la robótica colaborativa.
- CE15. Examinar las diferentes etapas que forman la cadena de valor del sector y sus mecanismos de control de calidad y evaluar las posibilidades de mejora de la eficiencia de sus procesos mediante la aplicación de metodologías habilitadoras de la transformación digital.
- CE16. Identificar, analizar e integrar las diferentes fuentes de información de datos generados en la empresa y aplicarlas al proceso de toma de decisiones.
- CE17. Identificar y analizar los procedimientos técnicos y administrativos necesarios para la elaboración y puesta en marcha de proyectos de transformación digital de empresas del sector.
- CE18. Analizar con espíritu crítico la evolución de la transformación digital dentro de la empresa para apoyar de forma creativa la innovación tecnológica.
- CE19. Planificar las diferentes etapas del desarrollo de proyectos en el ámbito de la transformación digital de empresas del sector, incluyendo el diseño, la redacción y firma, si fuera necesaria.

1.8 Estructura de la memoria del TFM

La memoria se desarrolla a lo largo de 7 capítulos diferenciados:

CAPÍTULO 1 – INTRODUCCIÓN ➔ constituye el punto de partida de la memoria, recogiendo todos los puntos principales de modo que, visualizando este capítulo, se consiga tener una perspectiva general de todo lo que se ha desarrollado en el trabajo.

CAPÍTULO 2 – FUNDAMENTOS TEÓRICOS ➔ este capítulo recoge todos los conocimientos teóricos que fundamentan los contenidos desarrollados a lo largo del trabajo.

CAPÍTULO 3 – ESPECIFICACIONES TÉCNICAS ➔ aquí se encuadran todas las especificaciones, tanto generales como específicas, que se han utilizado para el desarrollo del proyecto: viabilidad, métricas de estimación, normas, arquitectura, herramientas, etc.

CAPÍTULO 4 – METODOLOGÍA EXPERIMENTAL ➔ en este capítulo se especifican todos los pasos que se han seguido para el desarrollo del modelo propiamente dicho.

CAPÍTULO 5 – RESULTADOS Y DISCUSIÓN ➔ en este punto se exponen con detalle todos los resultados obtenidos y se ofrecen una serie de comentarios sobre los mismos con objetivo de analizar y aclarar cada uno de ellos.

CAPÍTULO 6 – CONCLUSIONES Y TRABAJOS FUTUROS ➔ aquí se presentan las conclusiones generales que se obtienen tras la realización del proyecto y se presentan una serie de posibles investigaciones y proyectos que se puedan abordar en el futuro.

CAPÍTULO 7 – BIBLIOGRAFÍA ➔ es el capítulo que cierra la memoria y recoge toda la bibliografía utilizada para la redacción de la misma.

CAPÍTULO 2 - FUNDAMENTOS TEÓRICOS

2.1 Transformación digital.

La transformación digital permite ganar competitividad en un mundo cada vez más digitalizado, como bien comenta un estudio del instituto tecnológico de Massachusetts [17], en el que se afirma que las empresas más tecnológicas son las más rentables y aportan más valor a sus clientes.

Es por eso que la definición de transformación digital es compleja, pero se puede decir que esta consiste en la integración de nuevas tecnologías en todas las áreas de una empresa para cambiar su forma de operar, con el objetivo de optimizar los procesos, mejorar su competitividad y ofrecer un valor superior a sus clientes.

Este proceso no consiste solo en comprar los equipos más actualizados y modernos e implantarlos en la empresa, sino que requiere un cambio de mentalidad de forma vertical en la corporación, desde los directivos hasta el último de los empleados, para conseguir afianzar con éxito el camino hacia la transformación digital [18].

Como todo proceso de innovación, la transformación digital ofrece una serie de ventajas:

- Impulsa la cultura de innovación en la empresa.
- Mejora la eficiencia de los procesos.
- Fomenta el trabajo colaborativo.
- Ofrece una capacidad de respuesta rápida a los cambios.
- Abre las puertas a nuevas oportunidades de negocio.
- Aporta valor y mejora la experiencia de los clientes.

Pero al igual que ventajas, esta presenta una serie de inconvenientes, siendo el principal de ellos el componente económico. Llevar a cabo un proceso de transformación digital corporativo supone una gran inversión, pues se requiere de personal cualificado para usar las nuevas tecnologías, así como los propios dispositivos que se han de implantar. En este caso hay dos caminos: o bien se generan procesos formativos para enseñar a los empleados que ya se tienen en plantilla, o se contratan nuevos empleados y se reorganiza al resto, siendo ambos costosos tanto en dinero como en tiempo. Además, es este tiempo otra de las principales desventajas de la implantación de procesos digitalizados en una empresa, pues no es una cosa que se consiga de un día para otro, si no que requiere un tiempo de evaluación, desarrollo y posterior implantación que puede llevar meses e incluso años si no se gestiona de forma correcta.

2.2 Ciencia de datos.

Los datos son los que dan paso a un sinfín de nuevas tecnologías y soluciones, pues es de ellos de donde se derivan los conocimientos empresariales más importantes [19].

Pero todo esto es imposible si no se tienen las herramientas necesarias para el análisis y el tratamiento de esos datos, siendo este un punto fundamental. Primero, porque hay que manejar una cantidad inmensa de datos que, sin los algoritmos y los profesionales adecuados, serían imposibles de gestionar; y segundo, porque los datos en bruto no

son de utilidad ninguna, si no que estos han de ser filtrados y preparados para poder hacer las lecturas correctas que se están buscando.

Este problema viene derivado del Big Data, pues este involucra datos cuyo volumen, variedad, velocidad y complejidad requieren nuevas técnicas, algoritmos y análisis para extraer conocimiento de ellos.

Es aquí donde entra en juego lo que se denomina como “Ciencia de datos”, que es la ciencia que se encarga de extraer dicho conocimiento de las cantidades masivas de datos existentes [20].

En la Figura 2.1 se pueden observar los cuatro ejes principales del Big Data [21]:

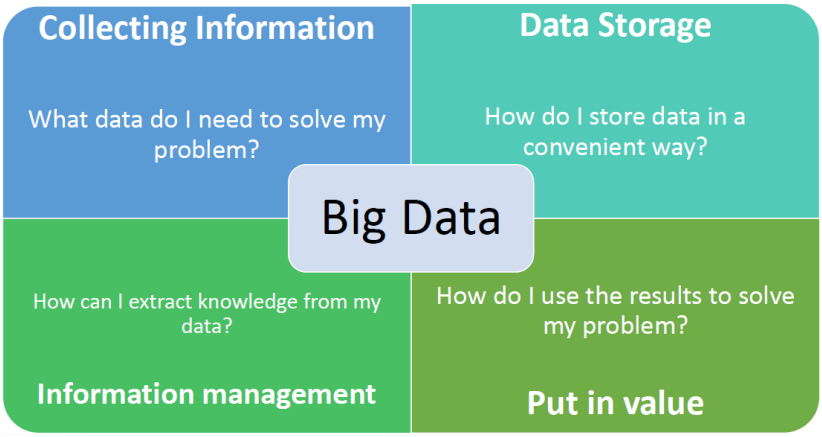

Figura 2.1. Ejes del BIG DATA [20].

- Recabar información ➔ ¿Qué datos se necesitan para resolver el problema?
- Almacenamiento de datos ➔ ¿Cómo se han de almacenar los datos?
- Gestión de la información ➔ ¿Cómo se consigue extraer conocimiento de los datos?
- Poner en valor ➔ ¿Cómo se han de usar los datos para solucionar el problema?

De esta forma se puede analizar el ciclo de vida de la ciencia de datos de la siguiente manera (Figura 2.2):

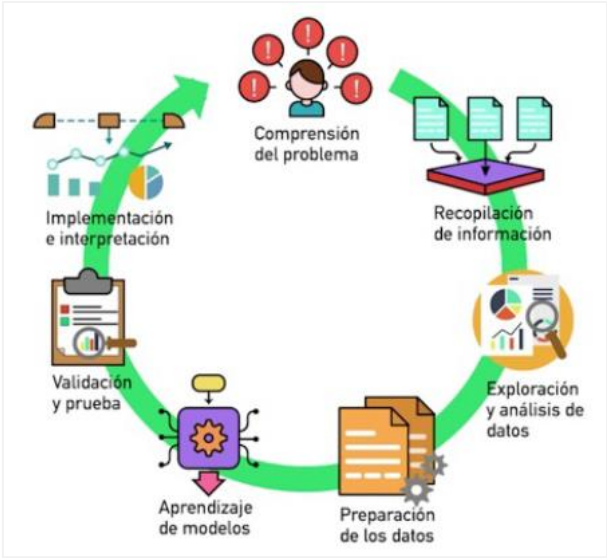

Figura 2.2. Ciclo de vida de la ciencia de datos [20].

- Entender el problema ➔ a través del conocimiento de los expertos, e identificar el objetivo a alcanzar.
- Recopilación de información ➔ tanta como sea posible. Esta información puede proceder de dos vertientes: datos que se han recogido por parte de la propia empresa (datos de encuestas, información de los clientes, datos logísticos, transacciones financieras, etc.), o datos públicos compartidos y distribuidos por cualquier persona libremente.
- Explorar y analizar los datos ➔ mediante el cálculo de estadísticas descriptivas o visualización.
- Preparación de datos ➔ filtrando y analizando sus características para conseguir obtener unos datos de calidad que ayuden a obtener información para, de igual forma, conseguir modelos de calidad. Esto se consigue a través de un proceso que se denomina curación de datos, el cual consiste en una gestión activa y continua de los datos a lo largo de su ciclo de vida. Además, este proceso comprende el mantenimiento y gestión de los datos, así como su almacenamiento en repositorios, de forma que resulten útiles para sus usuarios finales. Estos han de ser fácilmente recuperables para futuras investigaciones o cualquier otro uso.
- Modelar el aprendizaje ➔ usando técnicas de Machine Learning. Este comprende en general todo el proceso de curación, análisis, validación y visualización de resultados, previo a lo que comprende la extracción de conocimiento y obtención de conclusiones. Una vez usadas las técnicas de Machine Learning, los datos en bruto de la entrada se han transformado en datos ordenados y con un significado final.
- Validación ➔ pruebas para comprobar la calidad del modelo obtenido. Esta depende sobre todo de si el aprendizaje está dentro del grupo de supervisado o de no supervisado. En el aprendizaje supervisado es obligatorio validar el modelo, mientras que en el no supervisado no lo es en tanta medida. La explicación se encuentra en el siguiente punto.
- Implementación e interpretación ➔ a través de informes y gráficos que determinan si los modelo ofrecen información útil sobre el problema en cuestión.

2.3 Machine Learning.

El Machine Learning es el área de estudio encargada de diseñar algoritmos que son capaces de construir modelos automáticos a partir de una serie de datos de entrada [22].

El principal recurso para el aprendizaje en Machine Learning son los datos, usando como técnica un procedimiento muy similar al cognitivo en los seres vivos, basado en la práctica o la experiencia. Así, el modelo, a partir de los datos de entrada, los analiza, aprende y adquiere experiencia sobre ellos, para que, cuando le vuelvan a llegar otros datos, sea capaz de devolver el conocimiento adquirido previamente.

Como se puede ver en la Figura 2.3, existen cuatro tipos de enfoque en la realización de modelos en Machine Learning:

- Descriptivo ➔ para comprender qué ha sucedido.
- Predictivo ➔ para conocer qué va a suceder.
- Diagnóstico ➔ para determinar por qué ha sucedido.

- Preceptivo ➔ para conocer cómo se puede hacer que eso suceda.

Figura 2.3. Tipos de modelos de Machine Learning en función de su objetivo [22].

Pero la mejor división de los tipos de modelos existentes en Machine Learning es la que se observa en la Tabla 2.1:

Tabla 2.1. Agrupación de los tipos de modelos de Machine Learning [22].

	SUPERVISADO	NO SUPERVISADO
CATEGÓRICO	Clasificación	Reglas de asociación
CONTINUO	Regresión	Clustering

En breves palabras, el aprendizaje supervisado es aquel en el que los datos de entrada están vinculados a uno o varios datos de salida, mientras que, en el no supervisado, los datos de entrada no están ligado a ningún dato de salida.

De este modo, se pueden destacar las siguientes consideraciones acerca del Machine Learning:

- La etapa de aprendizaje generalmente se conoce como entrenamiento.
- Tiene la misma relación que la formación humana: ensayo y error.
- Una buena formación implica un buen modelo, por lo que, para un buen entrenamiento son fundamentales las siguientes cuestiones:
    - Datos de entrada de calidad.
    - Configuración de algoritmos adecuada (tiempos, parámetros, etc.).
- A mayor cantidad y calidad de los datos, más robusto y preciso será el modelo:
    - Por esto los algoritmos de Machine Learning se conocen como “Codiciosos de datos”.
    - El conocimiento está en los datos, pero es necesario extraerlo.
    - Cuanta más información se tenga sobre el problema, más patrones o reglas diferentes se podrán obtener.
    - Es imprescindible tener en cuenta la expresión “Basura que entra basura que sale”.
- El conjunto de entrenamiento debe ser lo suficientemente representativo del total de datos que conforman el problema a resolver.

- Se debe evitar el sesgo de datos, ya que el algoritmo procesa directamente toda la información de entrada.
- La etapa de exploración de datos permite a los profesionales garantizar la calidad de los datos.

Atendiendo ahora al grupo de modelos que conforman el aprendizaje supervisado, las consideraciones a tener en cuenta son:

- Es obligatorio determinar si el proceso de formación ha sido exitoso.
- Para hacerlo, un proceso de validación garantiza la corrección del modelo para datos nuevos o no vistos.
- El llamado conjunto de prueba se usa para este propósito, ya que contiene instancias del problema original que no pertenecen al conjunto de entrenamiento.
- En otras palabras, el conjunto de datos inicial debe dividirse en dos conjuntos disjuntos: entrenamiento y prueba.
- El modelo se disparará para cada conjunto de prueba, y si la predicción corresponde a la de la variable de salida, se puede determinar que el aprendizaje ha sido satisfactorio.

De igual modo, teniendo en cuenta el aprendizaje no supervisado, las consideraciones son:

- En este caso, un conjunto de prueba no es interesante en absoluto.
- Obviamente, como no hay una "variable de salida", no se puede determinar la bondad del modelo "comparando" con el valor original.
- Se necesita otro tipo de análisis sobre el conjunto de datos completo, por lo que existen diferentes medidas para estimar si el modelo obtenido se ajusta a los datos o no.
- Su cómputo depende del interés del usuario, y del tipo de paradigma de aprendizaje al que nos dirigimos (supervisado vs. no supervisado).
- En general, todas estas métricas tienen como objetivo determinar el error que comete el modelo al predecir la salida para nuevos datos de consulta de entrada.

En estas dos últimas listas de consideraciones se demuestra la importancia y la necesidad de validar los modelos que se generen, principalmente en aprendizaje supervisado, y el proceso es el siguiente: en aprendizaje supervisado existen dos fases muy bien diferenciadas que son el entrenamiento (*train*) y la validación (*test*). Se deben utilizar dos conjuntos de datos totalmente independientes. De utilizar algún ejemplo de test durante el entrenamiento, se estará sobreestimando la calidad del modelo generado. Así, los principales mecanismos de validación son:

- Retención o *Hold-Out*.
- Validación cruzada de k particiones o *k-fold cross validation*.
- “Dejar uno fuera” o *Leave one out validation (LOOV)*”

De igual modo, también hay que validar en cierto modo el aprendizaje no supervisado, pero, ¿cómo se puede contrastar si resultados obtenidos tienen la calidad suficiente?, pues no tiene sentido aplicar una validación cruzada, ni ningún tipo de validación basada en el conocimiento de ningún tipo de valores de salida. La pregunta a responder en este caso sería: ¿es la información que hemos descubierto fiable?, lo que sería equivalente

a preguntar: ¿Está la información que hemos descubierto soportada por los datos? ¿Existe fundamentación teórica y/o es lógico el resultado? Es así como se consigue evaluar si el modelo obtenido ofrece buenos resultados o no, siempre sujetos al punto de vista de un experto en estos datos.

A continuación, la Tabla 2.3 muestra un resumen de las ideas principales descritas en este punto.

Tabla 2.3. Ideas principales sobre el Machine Learning [22].

Sin análisis, los datos son solo "ruido"			
Es necesario un proceso automatizado para extraer conocimiento de grandes cantidades de datos	Confirmar las hipótesis, explicar los fenómenos y encontrar patrones desconocidos	Es necesario un conocimiento experto sobre el problema para entender los resultados	Elegir el método y las herramientas correctas es fundamental

2.4 Aprendizaje supervisado.

Como en este proyecto se usa un tipo de modelo que pertenece al grupo del aprendizaje supervisado, solo se atiende a la descripción de este tipo de aprendizaje.

Para la explicación concreta del aprendizaje supervisado, se pueden utilizar un par de ejemplos cotidianos, teniendo en cuenta que uno de los campos más habituales de aplicación del Machine Learning es el de la salud [22]:

- Herramienta de diagnóstico, para determinar la categoría a la que pertenece un paciente (sano o enfermo).
- Herramienta para ajustar el nivel de dosis concreto de un medicamento, en este caso calculando un valor numérico.

Está claro que, en ambos ejemplos, el objetivo es realizar una predicción del valor de una variable de salida, como el tipo de paciente, o la dosis del medicamento.

De este modo, comparando, el ser humano realiza esta tarea en base a su conocimiento basado en la experiencia con otros casos similares (pacientes y medicamentos), mientras que en el caso del Machine Learning, esta experiencia se recopila a partir de instancias de un conjunto de datos.

Por eso, los casos en los que se busca predecir un valor de salida en base a unos datos de entrada son los que se encuentran dentro del aprendizaje supervisado.

Las dos principales vertientes de este tipo de aprendizaje son la regresión y la clasificación.

2.4.1 Clasificación.

Por el mismo motivo que en el apartado anterior, en este caso solo de describe teóricamente el concepto de clasificación, ya que el modelo utilizado pertenece a este grupo de aprendizaje.

La clasificación se aborda cuando el problema presenta una serie de clases o etiquetas con respecto a la salida del modelo. Es decir, entra una transacción con una serie de

datos de entrada y el modelo tiene que decidir a qué etiqueta o a qué clase pertenece dicha transacción [23].

En este caso, lo que sucede es que entran los rechazos, marcados por una serie de características, y la salida correspondiente a cada uno de ellos es su forma de finalización, es decir, en firme o reprocesable.

El objetico de este método es obtener lo que se conoce como “discriminante”, que consiste en una función que es capaz de distinguir entra las clases de salida del modelo. De este modo, se genera una división (por ejemplo, una simple línea recta o curva) que parte el conjunto de datos en dos o más secciones, creando zonas que identifican a cada una de las clases (Figura 2.4).

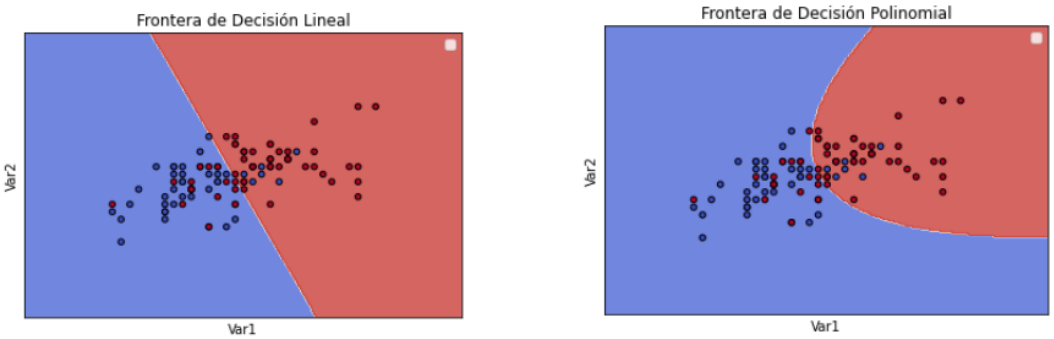

Figura 2.4. Representaciones de funciones discriminantes [24].

Las funciones discriminantes tienen una doble misión: actuar como herramientas predictivas, que es el caso hacia el que se enfoca este proyecto; y permitir la extracción de conocimiento de los datos mediante la propia representación del modelo.

Como viene siendo habitual, es fundamental evaluar la calidad del modelo una vez completado. Así, una de las principales formas de evaluación la componen la “Matriz de confusión”, la cual consiste en una matriz que registra verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). En la Figura 2.5 se puede ver el esquema de una matriz de confusión.

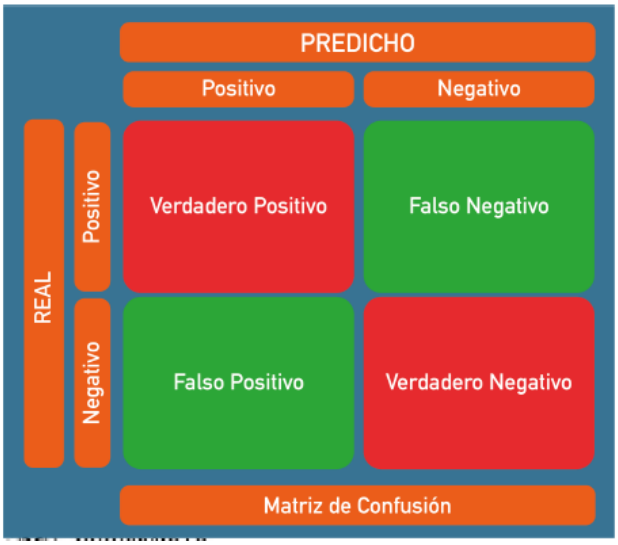

Figura 2.5. Esquema de una matriz de confusión [24].

Una de las principales ventajas de este tipo de aprendizaje es su alta interpretabilidad. Este consta de una representación simple y cercana a la cognición y semántica humana, que explica los datos de una manera parecida a como lo haría una persona, por lo que es muy sencillo de entender [25].

En general, existen dos grandes grupos en los que se pueden dividir los modelos de clasificación: modelos de clasificación de caja blanca, y modelos de clasificación avanzada (o caja negra) [24].

- Modelos de clasificación de caja blanca.

  Se consideran la primera línea de ataque cuando se quiere resolver un problema de clasificación. Esto se debe a que se obtienen de forma relativamente rápida y simple, y a que son en general altamente interpretables ya que se pueden examinar de forma sencilla las componentes del modelo e identificar las variables utilizadas para realizar la división entre las clases. Además, permiten al usuario comprender si éstas tienen un sentido o no con respecto al problema.

  Los principales tipos son:

  - Modelos de clasificación lineal ➔ se basan en una ecuación lineal, al igual que en regresión, en la que cada una de las variables de entrada constituye un peso diferente en la decisión de clasificar la variable de salida. Como ventajas se puede decir que son muy eficientes en su uso para una aproximación inicial al problema, son altamente interpretables y suelen funcionar bien en alta dimensionalidad. Aunque, por otro lado, hay que tener cuidado con variables correlacionadas. Además, la estimación resulta altamente sensible a los errores aleatorios en la variable de salida y, si los datos del problema son relativamente complejos, es recomendable utilizar técnicas no lineales más sofisticadas.
  - Modelo del vecino más cercano ➔ al igual que en regresión, la premisa para realizar la clasificación sobre una nueva instancia se basa en analizar la clase para instancias similares. El parámetro k (número de vecinos a tener en cuenta) es un parámetro crítico al que se le suele asociar siempre un número impar para evitar empates en la clasificación. Como ventajas, destacar que es muy eficaz y muy válido en clasificación, tanto como para regresión. Es un algoritmo rápido, pero no por eso ofrece menos capacidad predictiva, y está disponible en la mayoría de los paquetes de software. Pero su principal desventaja es que es muy ineficiente con respecto a la memoria de almacenamiento por lo que, a mayor número de instancias, más lenta es la predicción.
  - Árboles de decisión ➔ está basado en reglas de decisión simples, o condiciones, con el formato "si-entonces-si no" (*IF-THEN-ELSE*), normalmente dicotómicas (grupos de dos). Existe un orden jerárquico en la aplicación de las reglas, que se van encadenando hasta dar la decisión final. Cuánto más arriba del árbol, más decisiva ha de ser la variable que se cuestione. Como ventajas principales se encuentra la facilidad de usar y de interpretar qué presentan, además de que son mucho más fácilmente escalables que otro tipo de técnicas. En cuanto a las desventajas, no maneja variables de entrada de tipo numéricas y tiene dificultades para gestionar los valores perdidos. Además, puede insertar problemas de sobre aprendizaje debido a que no detecta con claridad las correlaciones existentes entre variables.

- Modelos de clasificación avanzada.

  Este tipo de modelos más complejos se suelen usar cuando prima la máxima capacidad predictiva de la variable de salida frente a una mayor interpretabilidad.

  Es cierto que su comportamiento es muy robusto frente a problemas difíciles, pero esto implica un mayor coste computacional en la fase de aprendizaje y, además, la necesidad de un mayor nivel de conocimiento del algoritmo para usarlo correctamente.

  Existen dos principales tipos de modelos de caja negra que son:

  - Máquinas de vectores soporte (SVM) ➔ esta es una de las herramientas preferidas de los científicos de datos ya que son capaces de obtener una alta precisión incluso en problemas complejos. Su objetivo es encontrar un hiperplano de separación entre las distancias de dos clases, teniendo como factor determinante lo que se denomina como "Coste (C)". Un valor bajo del coste aceptaría cometer un cierto número de errores de clasificación, bajando ligeramente la calidad de predicción obtenida en el conjunto de entrenamiento, pero buscando una mejor generalización en test. Mientras que un valor alto de este parámetro permite ajustar mejor el modelo sobre los datos de entrenamiento, pero implicaría un mayor riesgo de sobre aprendizaje.
    Como principales ventajas, se puede destacar que son muy eficaces en problemas con una alta dimensionalidad. Además, tienen un buen comportamiento cuando el número de variables es mayor que el número de instancias, adaptándose bien a problemas muy distintos entre sí. Como desventajas, es cierto que presentan cierta dificultad en la parametrización del algoritmo, no están adaptados para variables no numéricas y tienen exclusividad en problemas binarios, es decir, no permiten la clasificación en problemas de más de dos clases.
  - Random forest ➔ este método se basa en un principio llamado "Bagging (Boostrap aggregating)". Consiste en tomar M grupos de muestras, y entrenar M clasificadores diferentes sobre los subconjuntos anteriores. Para una nueva consulta, hay que hacer que todos los clasificadores lancen su predicción y tomar el resultado promedio o mayoritario. Si los clasificadores tienen errores independientes, entonces su unión puede mejorar el rendimiento. Dicho de otra manera: la varianza en la predicción se reduce ya que no afectan los errores aleatorios que un único clasificador puede cometer (Figura 2.6).

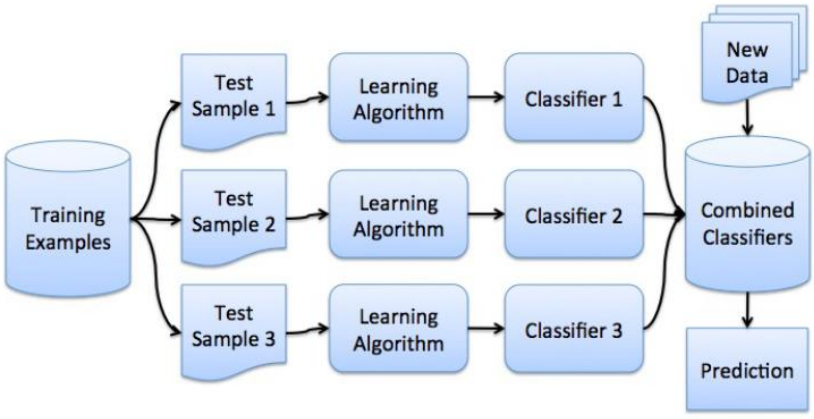

Figura 2.6. Esquema de funcionamiento del modelo Random Forest [24].

Pero, además, el random forest añade una nueva componente a esta metodología, que es la realización de una selección de variables, con lo que se obtiene una doble ganancia: por un lado, se incrementan las diferencias entre cada conjunto, y por tanto entre cada modelo generado; y por otro, el tiempo de aprendizaje y la complejidad de cada modelo se ve reducida

Este es el tipo de algoritmo que ha seleccionado la empresa para lanzar su modelo predictivo. En el Capítulo 4, se desarrollarán más especificaciones sobre el mismo, ya más enfocadas en la aplicación práctica que ocupa a este proyecto.

En general, en este capítulo se han descrito todas las nociones teóricas necesarias para comprender el problema que se presenta frente al proyecto, así como las distintas soluciones que se han ido desarrollando para resolverlo.

CAPÍTULO 3 - ESPECIFICACIONES TÉCNICAS

En la siguiente Tabla 3.1 se presentan las distintas especificaciones que se deben tener en cuenta para el desarrollo completo del proyecto en relación con: aprobaciones, equipos, herramientas, manuales, desarrollo, pruebas e implantación. Cada una de las fases se describe de forma más precisa en sus correspondientes apartados.

Tabla 3.1. Tabla de especificaciones técnicas.

ESPECIFICACIONES TÉCNICAS	
Aprobaciones	> Presentación del proyecto al comité de financiación > Presentación del proyecto al equipo de sistemas - Equipo desarrollador > Aprobación económica > Aprobación técnica por el equipo de sistemas
Equipos	> Ordenador portátil MSI intel core i5 > Router - Conexión a la red vía WIFI
Herramientas	> KNIME > Excel > Powert Point > Word > Adobe PDF Reader > Salesforce > App Modelo predictivo de rechazos de Endesa
Manuales	KNIME > KNIME Analytics Platform Installation Guide > KNIME Workbench Guide > KNIME Best Practices Guide > Extensions and Integrations Guide > KNIME Flow Control Guide > KNIME Components Guide > KNIME Integrated Deployment Guide > KNIME File Handling Guide > KNIME Workflow Invocation Guide RECHAZOS > Manual cortes y bajas impago v1 > Manual de operativa gestión rechazos electricidad b2c_v30 > Manual de operativa gestion rechazos formato r1 v9-cosmos > Manual de operativa gestion rechazos gas b2c_v23 > Manual Productos Happy sin DH con TG para paso a DH V5 > Notas Operativas Rechazos Electricidad B2C 04032022 > Notas Operativas Rechazos Electricidad B2C 04032022

Tabla 3.1. Tabla de especificaciones técnicas (continuación).

Desarrollo	> Extracción de los datos > depuración de los datos > Generación del flujo en KNIME > Comparación de modelos > Lanzamiento del modelo propio > Comprobación > Ajuste > Obtención del modelo final
Pruebas	> Ejecución del modelo en el entorno de pruebas de Salesforce > Visualización de resultados obtenidos > Comprobación del cumplimiento de todas las funciones requeridas > OK a las pruebas al pasar todos los filtros
Lanzamiento/Implantación	> Pasado el periodo de prueba se instala en el de producción > Seguimiento continuo de las operaciones que realice el modelo de forma automática > Reporte de informes de seguimiento

CAPÍTULO 4 - METODOLOGÍA EXPERIMENTAL

4.1 Obtención de los datos.

Los datos se han extraído del propio sistema operativo de la compañía, pues es importante trabajar con datos reales para que el proyecto funcione con la mayor cercanía posible a los casos de los verdaderos clientes de la empresa.

En este caso, los datos utilizados como datos para entrenar el modelo son todos los rechazos finalizados en este año 2022, desde el día 1 de enero al 7 de junio, entendiéndose como finalizados los que ya se resolvieron, bien como en firme o bien como reprocesable. Con ello se han conseguido una totalidad de 52619 instancias una vez pasados por los filtros que se describen en el siguiente punto.

Los campos que caracterizan los rechazos cargados son: rechazo, pedido, empresa titular, tipo cliente, línea de negocio, territorio, distribuidora electricidad, distribuidora gas, CUPS, provincia, número días de parada, días de gestión, modo nueva solicitud, veredicto, motivo parada, explicación, producto, tipo de gestión de rechazo, envío carta, gestión telefónica, oferta original, tipo de solicitud, subtipo de solicitud, canal de entrada, solicitud ATR (Acceso de Terceros a la Red), fecha envío, fecha real envío, fecha de entrada, fecha fin actuación, creado por, cliente, NIF (Número de Identificación Fiscal) cliente, teléfono, motivo BOC, motivo de rechazo electricidad, motivo de rechazo gas, resto motivos rechazos, tipo carta, última modificación por, estado rechazo, fecha primera parada, nombre del propietario, id de la necesidad, necesidad, reiterado, rechazo pendiente de CIE (Certificado de Instalación Eléctrica) y tipo rechazo.

4.2 Depuración.

No todos los rechazos finalizados son válidos para introducirlos en el modelo, puesto que estos, aparte de haber finalizado, tienen que cumplir una serie de características.

La primera de ellas es que tiene que pertenecer al grupo del cost to serve (coste para servicio). Dentro de la propia comercializadora, Endesa está dividida en dos empresas diferentes: Energía XXI, que es la que gestiona el mercado regulado, y Endesa Energía, que es la que gestiona el mercado libre. Dicho esto, el cost to serve engloba a los rechazos que surgen debido a un alta de cliente desde la empresa de Energía XXI, y además, todos los que surgen de modificaciones y bajas, tanto los de Energía XXI como los de Endesa Energía.

Otro de los requisitos que han de cumplir los rechazos es que no sean reiterados, es decir, que la solicitud correspondiente no haya sido rechazada por segunda vez o más. Está estipulado que este tipo de rechazos han de ser gestionados manualmente por el equipo.

Y la última característica que han de cumplir es que, por supuesto, no hayan sido finalizados de forma automática. Ya se ha comentado con anterioridad que existen algunas paramétricas que hacen que determinados tipos de rechazos se resuelvan automáticamente. Si se solicitan al sistema todos los rechazos finalizados en un determinado intervalo de tiempo, también devolverá los que se hayan finalizado de

forma automática, por lo que estos se han de eliminar ya que no tiene sentido que se introduzcan en el modelo.

Esto se consigue a través de la inserción de una serie de filtros en el configurador de extracciones, de modo que el sistema devuelve solo los datos que realmente se necesitan.

4.3 Descripción de la herramienta *KNIME.*

KNIME es una herramienta gratuita de código abierto para crear y producir ciencia de datos utilizando un entorno fácil e intuitivo, por lo que no hace falta tener conocimientos profundos de programación para lanzar o generar modelos de Machine Learning [26].

Este software es tan sencillo de usar porque, para crear cualquier tipo de modelo de aprendizaje automático, solo hay que confeccionar un flujo en el que se conectan una serie de nodos determinados (Figura 4.1).

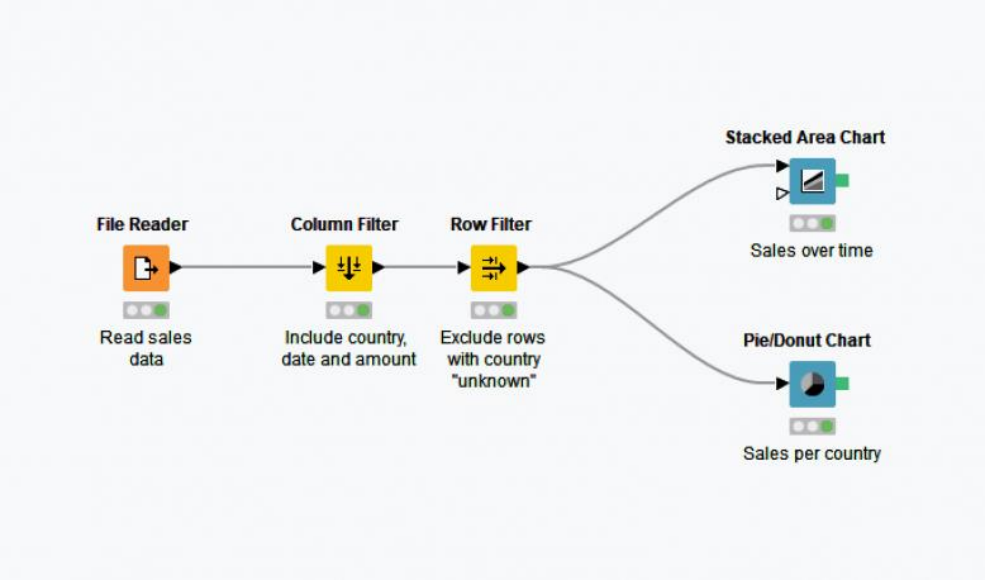

Figura 4.1. Ejemplo de flujo de trabajo en KNIME [26].

Los nodos pueden realizar todo tipo de tareas, incluida la lectura y escritura de archivos, la transformación de datos, la generación de modelos a partir de los conjuntos de datos, la creación de visualizaciones, etc. Un nodo se representa visualmente en *KNIME* como un pequeño icono compuesto de puertos de entrada y salida, así como del estado actual del nodo. El puerto de entrada se encuentra situado a la izquierda del nodo, el puerto de salida a la derecha del nodo y el estado en la parte inferior del nodo.

Las entradas son los datos que procesa el nodo, y las salidas son los conjuntos de datos resultantes. *KNIME* contempla distintos tipos de puertos de entrada o salida como mecanismos de transmisión de datos. Éstos suelen representar distintos tipos de conexiones de datos, entre los que se encuentran principalmente: datos, bases de datos y modelos.

Además, estos pueden tener distintos estados que se representa por un semáforo de colores: el rojo representa no configurado, el amarillo representa configurado, pero no

ejecutado, el verde representa ejecutado y un círculo rojo con una cruz, error en la configuración del nodo.

En la Figura 4.2 se representan las distintas partes de un nodo.

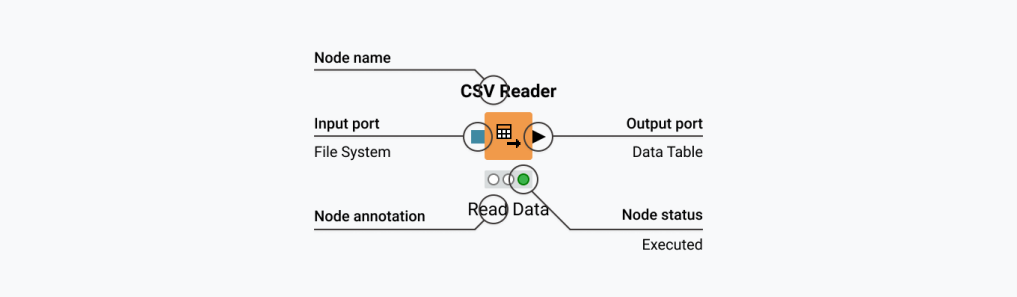

Figura 4.2. Partes de los nodos en KNIME [26].

A su vez, la interfaz se reparte de forma clara e intuitiva. En la parte izquierda, arriba se encuentra el explorador de archivos donde se pueden encontrar todos los flujos de trabajo creados en el dispositivo local, así como algunos de ejemplo. En el centro de la zona izquierda, se encuentran una serie de nodos que la propia herramienta reconoce que podrían ser de utilidad para el usuario en función de lo que quiere conseguir. Y en la parte más baja, está la librería de todos los nodos y funciones que posee la herramienta.

Por otro lado, atendiendo a la zona central de la pantalla, se encuentra un hueco grande en el que el usuario deberá desarrollar el propio flujo de trabajo, mientras que en la parte baja, la herramienta dispone de una pantalla dinámica que muestra las distintas variables que va procesando el flujo a medida que se va implementando.

Por último, en la zona de la derecha, existe una ventana en la que se explica el contenido y las funciones de los distintos nodos que se utilizan o se quieran utilizar.

En la Figura 4.3 siguiente se puede ver una representación de las distintas zonas explicadas.

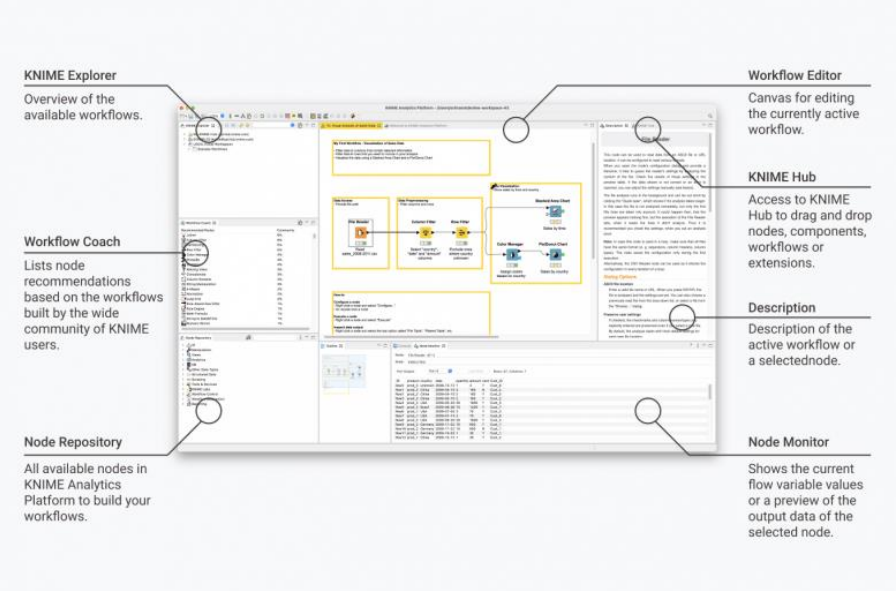

Figura 4.3. División del interfaz de KNIME [26].

4.4 Comparación de tipos de modelos.

En este punto se describe el flujo de trabajo diseñado para comparar la calidad de predicción de distintos tipos de modelos predictivos con el objetivo principal de justificar qué tipo de modelo será el más correcto para la realización de dicha predicción.

En la Figura 4.5 se puede observar el flujo completo implementado en *KNIME* para realizar la comparación.

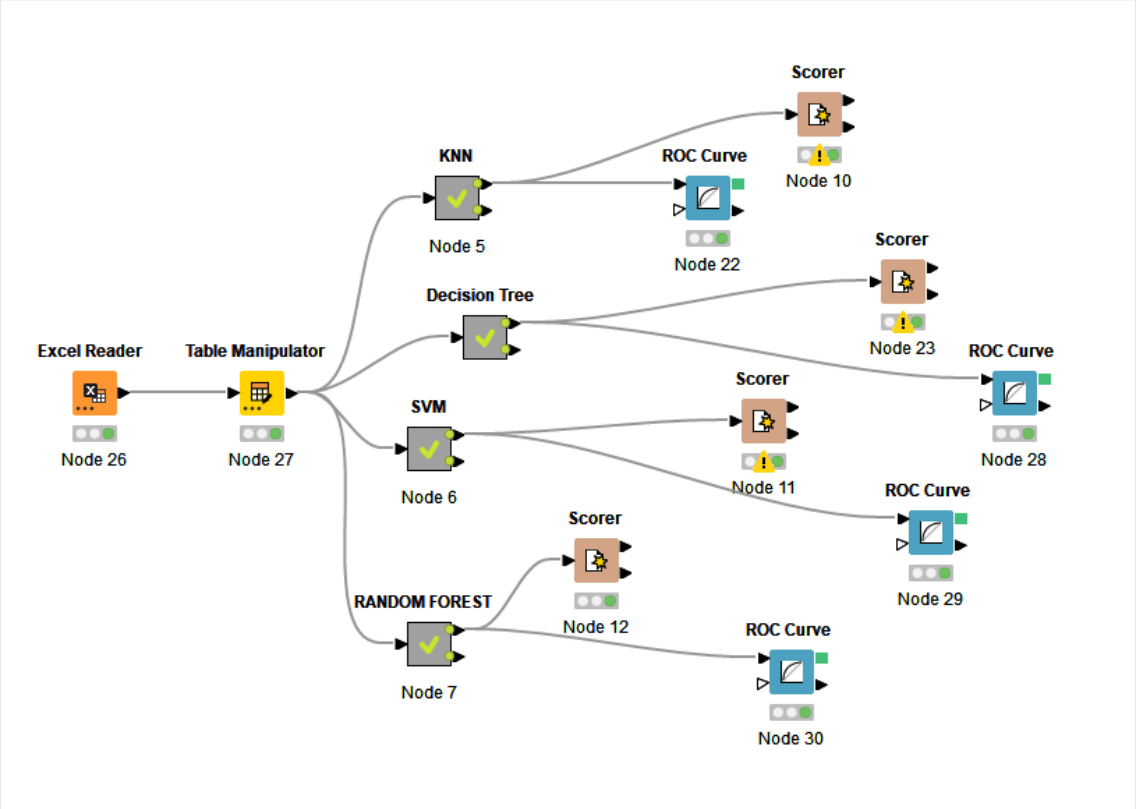

Figura 4.5. Flujo implementado para la comparación de modelos

Se ha realizado la comparación de cuatro tipos de modelos diferentes: KNN, Decision Tree, SVM, y Random Forest.

Los nodos utilizados han sido:

- Excel Reader ➔ este nodo está configurado para recibir una hoja de Excel con sus determinadas columnas y filas y convertirla en una tabla de datos interpretables por otros nodos del programa. En este caso, se utilizaron los rechazos ya finalizados en este año 2022, los mismos que se describen en el punto 4.1.1.
- Table manipulator ➔ permite modificar los parámetros de las columnas de la tabla que se le ofrezca en la entrada, en este caso, la que viene generada por el nodo anterior. Así, en este nodo se pueden cambiar los tipos de datos, por ejemplo, de string a integer, eliminar o añadir columnas, y algunas modificaciones más, con el objetivo de ajustar la tabla que recibe al gusto del operador. En este solo se introdujo el nodo por si era necesario hacer ajustes en algún momento, pero no se ha necesitado cambiar ningún parámetro de la tabla original.
- Cross Validation ➔ componen el siguiente nivel del flujo, donde se desarrollan propiamente los cuatro modelos que se han comparado. Cross Validation es un

metanodo, es decir, un nodo que contiene más nodos en su interior. En ellos se realiza lo que se conoce como validación cruzada. Es decir, se configura un número de divisiones en las que se van a agrupar los datos de entrada, en este caso 10. De este modo, el nodo divide los datos en diez grupos y lanza el modelo 10 veces. En la primera, entrena con los nueve primeros grupos y valida con el grupo número diez. En la segunda pasada, entrena con los ocho primeros más el último y valida con el grupo número nueve. Y así sucesivamente hasta que realiza las diez pasadas. Esto le permite ir ajustando los errores de una pasada a otra hasta que consigue el mejor resultado posible con la combinación de todas las comprobaciones.
En la Figura 4.6. se puede ver el interior del metanodo del Random Forest. El resto son iguales, solo que cambian los nodos Learner y Predictor por los de los otros tipos de modelos.

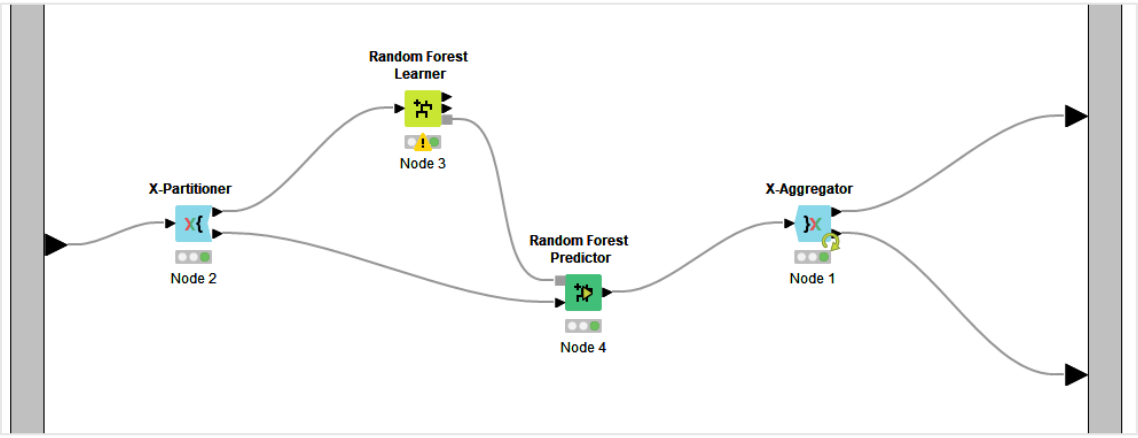

Figura 4.6. Metanodo Cross Validation del Random Forest

El nodo X-Partitioner es el que divide los datos en diez grupos. El Random Forest Learner es el que entrena y aprende de los datos de entrada y sus respectivas soluciones, mientras que el Random Forest Predictor es el que realiza el test con el grupo de validación que se citaba anteriormente. Una vez hecho esto, el X-Agregator une todos los resultados y los devuelve por los puertos de salida del metanodo.

- Scorer ➔ este es uno de los dos nodos que se encuentran tras la validación cruzada. El nodo Scorer se encarga de comparar los valores de la variable a predecir que están asociados a los datos de entrada, es decir, si el rechazo se había resuelto en firme o reprocesable, con la solución del rechazo que ofrece el modelo tras su aprendizaje y validación. De este modo, se pueden contabilizar los aciertos del modelo ya que se conocen cómo se resolvieron realmente estos rechazos, y así se puede ofrecer un índice de calidad de predicción.
- Roc Curve ➔ este nodo es el que se encarga de ofrecer el índice de calidad que se ha comentado en el punto anterior. Conforme más se acerque a 1, mayor será la calidad de predicción del modelo en cuestión.

En el siguiente capítulo se presentan los resultados obtenidos junto con algunos comentarios sobre ellos.

4.5 Implementación del Random Forest.

El desarrollo propiamente dicho del modelo predictivo que surge de este proyecto está agrupado en dos flujos diferenciados.

En la Figura 4.7 se puede observar el primero de ellos. Este consta simplemente de tres nodos que se encargan de la parte del aprendizaje del modelo.

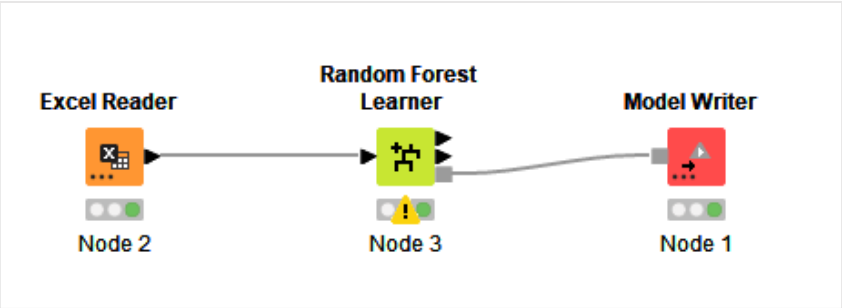

Figura 4.7. Flujo para la parte de aprendizaje del modelo

El primer nodo, Excel Reader, recibe una hoja de Excel con los datos que se han usado para el entrenamiento, en este caso, los rechazos finalizados en el año 2022 hasta el día siete de junio, que son los mismos que se describen en el apartado 4.1.1.

El siguiente nodo es el que se encarga del aprendizaje en base a los datos de entrada, en este caso del tipo de modelo Random Forest, que ha sido el seleccionado entre los cuatro comparados, al igual que el que seleccionó la propia empresa para el desarrollo de su modelo por motivos que se explican en el siguiente capítulo.

Habiendo ya aprendido de los datos de entrada, se obtiene un modelo propiamente dicho que se envía al último nodo, Model Writer. Este simplemente se encarga de recoger el modelo en su puerto de entrada y exportarlo a la dirección que se le indique en el equipo local. Esto se usa para guardar el modelo localmente y así poder usarlo cada vez que sea necesario en cualquier otro flujo de trabajo.

Es por esto último, que el punto de partida del segundo de los flujos que contiene el modelo comienza con el nodo Model Reader (Figura 4.8). A este primer nodo se le indica la dirección local del archivo del modelo citado anteriormente para que este nuevo flujo sea capaz de tener en cuenta el aprendizaje que se ha realizado en el flujo anterior.

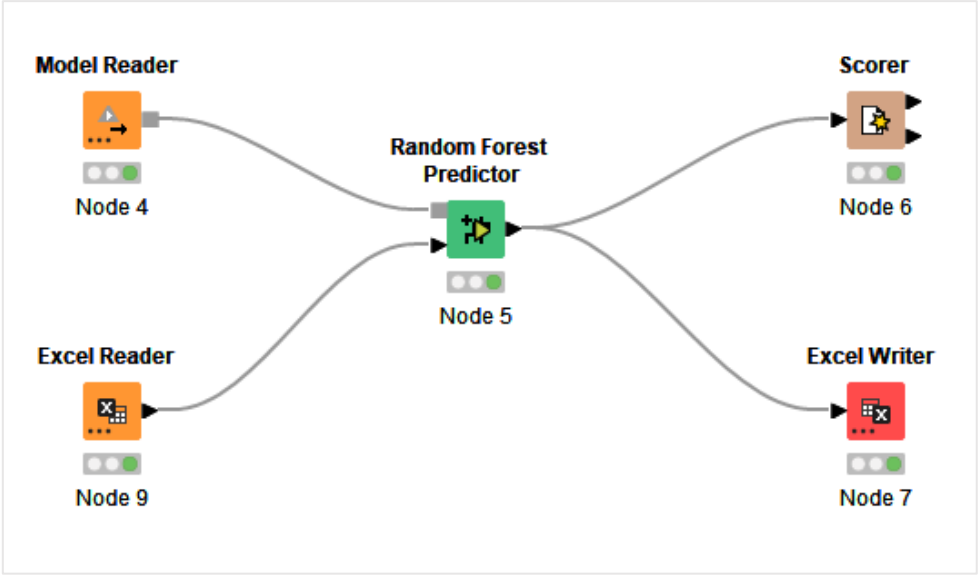

Figura 4.8. Flujo para el lanzamiento del modelo

Dicho esto, a continuación, se introducen nuevos datos y se comprueba si el modelo es capaz de predecir cuál será el valor de la variable de salida.

Para la introducción de los datos se vuelve a usar el nodo Excel Reader, aunque, en este caso, se introducen datos diferentes a los que se introdujeron en el flujo de aprendizaje, pues no tendría sentido que el modelo ofrezca una solución para los datos con los que ha sido entrenado.

Teniendo en cuenta esto, a este nodo se le ha introducido una hoja Excel que contiene los rechazos finalizados desde el día 8 de junio de 2022, hasta el día 20 de junio de 2022, un total de 2868 rechazos. Todos ellos se encuentran con una variable de salida asociada, es decir, se indica si se resolvieron en firme o reprocesable. Este dato será el que se utilice para valorar la calidad de la predicción una vez que el modelo ofrezca sus propios resultados.

El siguiente nodo es el Random Forest Predictor, el encargado de realizar la predicción sobre los nuevos rechazos introducidos y en base a los datos del modelo de aprendizaje.

Para terminar, se han introducido dos nodos más, el Scorer para valorar la calidad de la predicción y el Excel Writer para exportar una tabla Excel que revele los datos de salida del modelo y así poder visualizar correctamente los resultados, los cuales se exponen en el siguiente capítulo.

CAPÍTULO 5 - RESULTADOS Y DISCUSIÓN

5.1 Resultados de la comparación de modelos.

En este punto se van a analizar y comparar los resultados obtenidos en los índices de calidad de los modelos comparados.

5.1.1 Resumen de resultados, valoración y elección.

La Tabla 5.1 recoge todos los resultados con el objetivo de ofrecerlos agrupados para que sea más sencilla su comparación, a la hora de valorar cuál es el modelo que mejor resuelve el problema en cuestión y que será analizada en los siguientes apartados.

Tabla 5.1. Comparación de resultados de modelos predictivos

MODELO	KNN		DECISION TREE		SVM		RANDOM FOREST	
Rechazos de Entrada	52619		52619		52619		52619	
Rechazos Resueltos por el modelo	9122		51643		9122		52618	
	En Firme	**Reprocesable**	**En firme**	**Reprocesable**	**En firme**	**Reprocesable**	**En firme**	**Reprocesable**
	4508	4614	16019	35624	4508	4614	16317	36301
Aciertos Modelo	5099		50101		0		52083	
	En firme	**Reprocesable**	**En firme**	**Reprocesable**	**En firme**	**Reprocesable**	**En firme**	**Reprocesable**
	2530	2569	15032	35069	0	0	16133	35950
Fallos Modelo	4023		1542		9122		535	
	En firme	**Reprocesable**	**En firme**	**Reprocesable**	**En firme**	**Reprocesable**	**En firme**	**Reprocesable**
	1978	2045	987	555	4508	4614	184	351
% de Acierto	55,90%		97,01%		0,00%		98,98%	
Rechazos Perdidos	43497		976		43497		1	
% de Perdidos	82,66%		1,85%		82,66%		0,00%	

Los campos que se recogen para cada modelo en la tabla se describen de la siguiente manera:

- Rechazos de entrada ➔ muestra de rechazos de entrada que se introducen en los modelos para que realicen la predicción. Estos son los que se describen en el punto 4.1.1. Los datos son los mismos para todos los modelos para así poder realizar la comparación de forma correcta.
- Rechazos resueltos por el modelo ➔ este campo indica el número de rechazos de entrada para los cuales el modelo ha sido capaz de ofrecer un resultado, bien en firme o reprocesable, sin valorar aún si acierta o no. Eso se contempla porque hay rechazos que algunos modelos no pueden predecir, debido a los datos que los definen, mientras que otros tipos de modelos sí que pueden.
  Esto se refleja en la tabla pues, para los mismos rechazos de entrada, el modelo Decision Tree y el Random Forest son capaces de predecir casi la totalidad, mientras que los otros modelos no. Esto suele deberse al hecho de que existen campos vacíos, es decir, que alguno de los valores que caracterizan a los rechazos no están rellenos. Esta situación supone para los modelos KNN y SVM un factor determinante para la resolución del rechazo, por lo que, al estar vacíos, directamente los descarta y no ofrece resultado. Por el contrario, los otros dos tipos de modelos son menos restrictivos en ese sentido. Para estos, no es tan importante que alguno de los campos se encuentre vacío, pues tiene la capacidad de predecir basándose en los demás, lo que les da la oportunidad de ofrecer resultados para muchos más rechazos de entrada.

- Aciertos Modelo ➔ este parámetro se ha definido porque, aunque el modelo sea capaz de ofrecer un resultado, eso no significa que vaya a acertar. Este número indica la cantidad de rechazos resueltos por el modelo para los que este ha acertado en su predicción. Además, se desglosan en los dos tipos de resoluciones posibles.
- Fallos Modelo ➔ del mismo modo que el campo anterior, este número indica la cantidad de rechazos resueltos por el modelo para los que el este no ha acertado en su predicción. Además, se desglosan en los dos tipos de resoluciones posibles.
- % de Acierto = Aciertos modelo / Rechazos Resueltos por el modelo
- Rechazos perdidos = Rechazos de entrada – Rechazos resueltos ➔ muestra la cantidad de rechazos de entrada para los que el modelo no ha sido capaz de devolver un resultado.
- % de Perdidos = Rechazos perdidos / Rechazos de Entrada

Atendiendo a estas características, se valoran positivamente los modelos que ofrecen un mayor índice de acierto frente a un menor índice de perdidos.

Atendiendo a estas condiciones, los dos modelos que peor predicen son el KNN y el SVM. El KNN ofrece un 55,90% de acierto, lo que supone un valor bajo para este índice, el cual debería situarse al menos sobre un 85%. Además, presenta un 82,66% de rechazos perdidos, siendo esto inadmisible, ya que cuando se pierde un rechazo se debe a que, para los datos de entrada que se le ofrecen al modelo, este no es capaz de ofrecer ninguna solución. Pero aún peor es el resultado del SVM, el cual no acierta ningún caso de los predichos, ofreciendo también un porcentaje de perdidos del 82,66%.

Por el contrario, los resultados del Decision Tree y del Random Forest son muy positivos. Ambos ofrecen un porcentaje de acierto por encima del 97% y un porcentaje de perdidos por debajo del 2%, siendo el mejor de ellos el Random Forest, con un 98,08% de acierto y un 0% de perdidos.

Es por estos motivos por los que se ha decidido utilizar el Random Forest frente a los otros tipos de modelos, justificando así también la decisión que tomó Endesa en su momento a la hora de lanzar su modelo predictivo.

A continuación, en los siguientes puntos, se describen con más detalle los resultados ofrecidos por cada modelo.

5.1.2 Resultados del modelo KNN.

En la Figura 5.1 se pueden observar la matriz de confusión obtenida, además de los índices de calidad que ofrece la curva Roc.

De un total de 4508 rechazos que se resolvieron en firme, el modelo ha acertado solo en 2530. Del mismo modo, de los 4614 que se resolvieron como reprocesable, el modelo ha acertado en 2569. En general son unos valores muy bajos, lo cual se ve reflejado en la curva Roc. Hay que destacar que la línea azul es la que representa el resultado que se ha tomado como positivo, que es el de en firme, por lo que mientras más se acerque a 1 mejor. Por otro lado, la línea naranja representa el resultado que se ha tomado con valor negativo, que es el reprocesable, por lo que mientras más se acerque a 0 mejor. La línea central marca el valor de 0,5, el cual es el peor de todos para ambos casos.

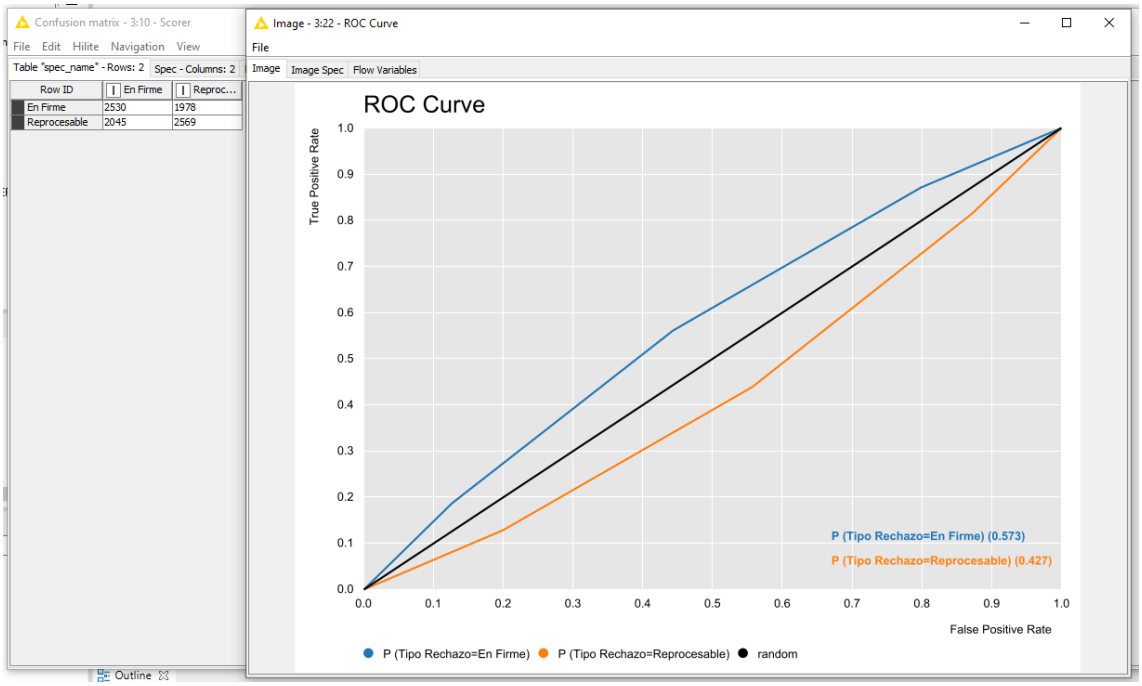

Figura 5.1. Resultados del modelo del tipo KNN

Hay que destacar que existen muchos rechazos que se han quedado por el camino sin predecir. En total, se han predicho 9122 rechazos (4508 en firme + 2614 reprocesables) frente a los 52619 que se ofrecieron de entrada. Son muchos los rechazos para los que el modelo no ha conseguido obtener una solución. Tras investigar, se ha llegado a la conclusión de que esto se debe a que, para modelos del tipo KNN, es importante que todos los campos que caracterizan a los rechazos vengan rellenos, es decir, que no haya en los rechazos algún campo vacío. El hecho de que existan valores perdidos se da en la mayoría de los rechazos, pues no es necesario que todos los campos vengan completos ya que estos dependen de las características de cada caso. Este dato de la gran cantidad de rechazos sin resolver comprende una valoración negativa para el KNN, lo que significa un punto muy importante que hace que se haya que descartar este tipo de modelo predictivo como solución del problema planteado.

Tras los resultados de la matriz de confusión, en la curva roc, el índice del en firme está en 0,573 y el del reprocesable se encuentra en 0,427. Ambos valores son muy bajos.

5.1.3 Resultados del modelo Decision Tree.

El siguiente tipo de modelo es el Decision Tree, cuyos resultados se pueden observar en la Figura 5.2.

La matriz de confusión muestra que, de 16019 rechazos que se resolvieron en firme, el modelo ha acertado en 15032. Del mismo modo, de los 35624 que se resolvieron como reprocesable, el modelo ha acertado en 35069. Ambos son muy buenos resultados, lo cual se demuestra en la matriz de confusión, siendo el índice de los en firme de 0,966, y el de los reprocesables de 0,036.

En este caso no sucede el problema citado anteriormente, en el que se daban una gran cantidad de rechazos sin resolver, ya que, de los 52619 rechazos de entrada, se han

resuelto un total de 51643 (16019 en firme + 35624 reprocesables), siendo este un buen resultado en cuanto a la cantidad de rechazos resuelto frente a los rechazos de entrada.

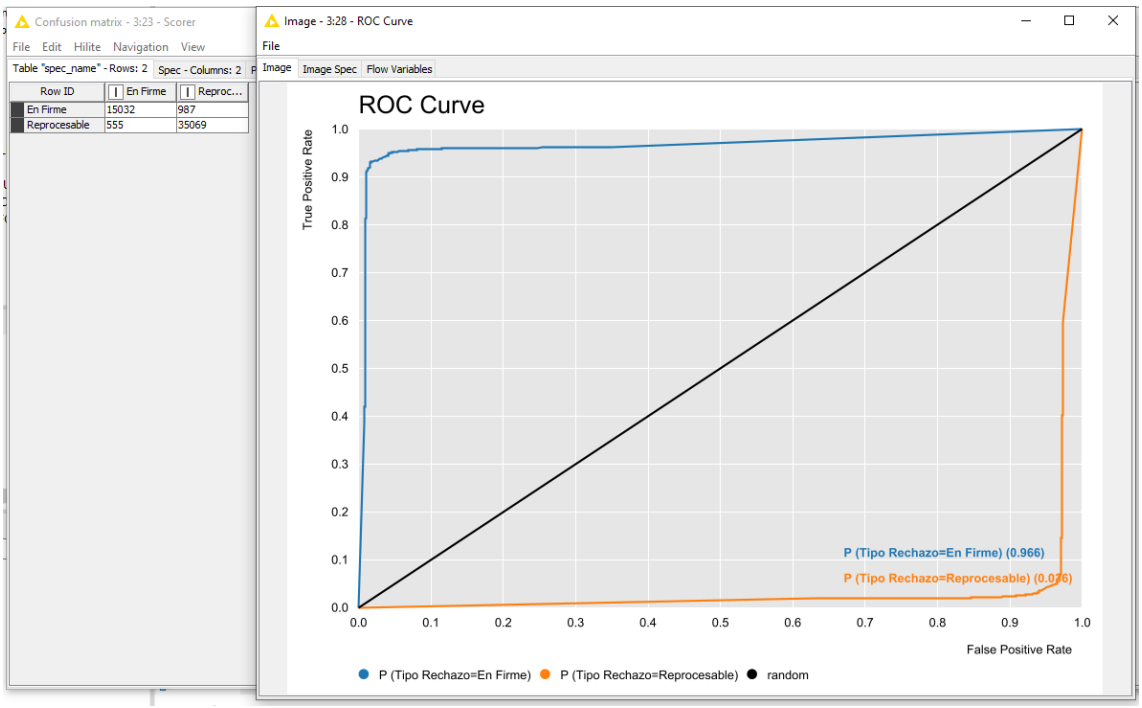

Figura 5.2. Resultados del modelo del tipo Decision Tree

5.1.4 Resultados del modelo SVM.

Otro tipo de modelo comparado es el SVM, o Máquina de Soporte Vectorial. En la Figura 5.3 se pueden observar los resultados obtenidos.

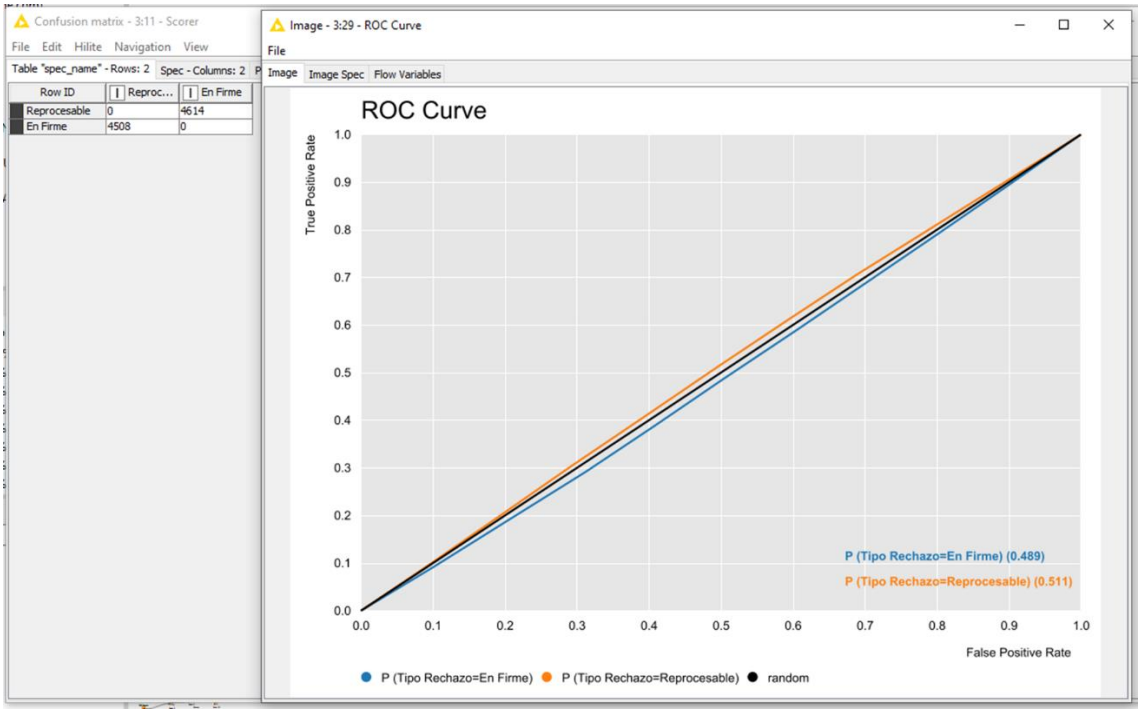

Figura 5.3. Resultados del modelo del tipo SVM.

Se puede deducir que el modelo no es un buen predictor de este problema, pues las líneas de la curva Roc están muy próximas a la línea central. Esto se ve reflejado del mismo modo en la matriz de confusión, donde se puede observar que no acierta en ninguna de las predicciones realizadas, 0 aciertos de los 4508 que se resolvieron en firme, y 0 aciertos de los 4614 que se resolvieron como reprocesables.

Además, vuelve a suceder el problema de la existencia la existencia de una gran cantidad de rechazos de entrada para los que el modelo no es capaz de ofrecer un resultado, al igual que en el KNN, pues de los 52619 rechazos de entrada, se han resuelto un total de 9122 (4508 en firme + 2614 reprocesables). En general, debido a estos resultados, el SVM no es una buena elección para resolver la predicción de los rechazos.

5.1.5 Resultados del modelo Random Forest.

Por último, se ha analizado el caso del Random Forest, que es el tipo de modelo que eligió la empresa para lanzar su modelo predictivo, como bien se justifica en el punto 5.1.1.

Se puede observar en la Figura 5.4 que los índices de calidad de predicción son muy positivos, o más bien excelentes. De un total de 16317 rechazos que se resolvieron en firme, el modelo ha acertado en 16133, fallando tan solo en 184. De igual modo, de los 36301 que se resolvieron como reprocesables, ha acertado un total de 35950, fallando solo en 351. Además, es el modelo que más se acerca a un cien por cien de rechazos predichos, ya que, de los de los 52619 rechazos de entrada, se han resuelto un total de 51618 (1637 en firme + 36301 reprocesables).

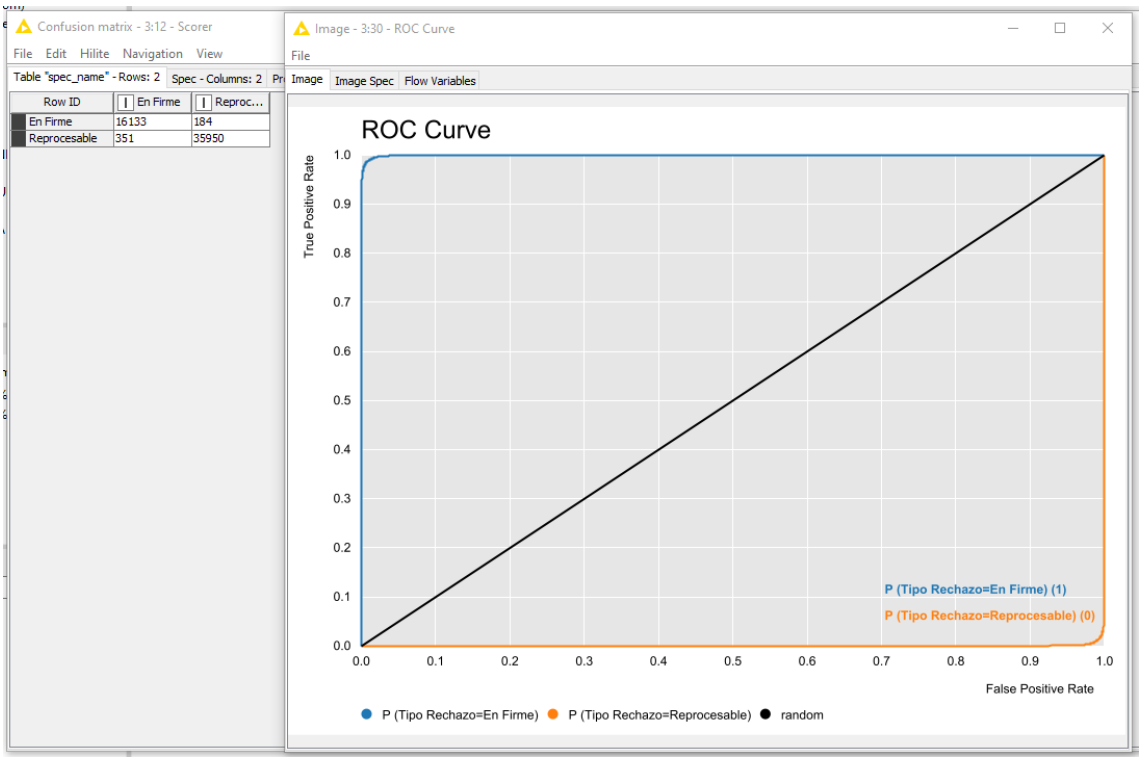

Figura 5.4. Resultados del modelo del tipo Random Forest.

Este elevado índice de acierto se ve reflejado en la curva Roc, donde ofrece un índice de 1 para los en firme, y un índice de 0 para los reprocesables, es decir, acierta siempre. De todas formas, se ha visto que esto último no es del todo cierto, ya que existen caoss en los que falla, pero es tan mínimo el porcentaje de error que comete el modelo que, para las tres cifras decimales del índice de calidad que ofrece la curva Roc, la aproximación se va al 1 y al 0. De nuevo se reitera el motivo de elección de este tipo de modelo, tanto en la empresa como en este mismo proyecto.

5.2 Resultados del modelo predictivo elegido.

Para analizar la calidad de predicción del modelo desarrollado en este proyecto, se ha vuelto a usar la matriz de confusión, la cual se muestra en la Figura 5.5.

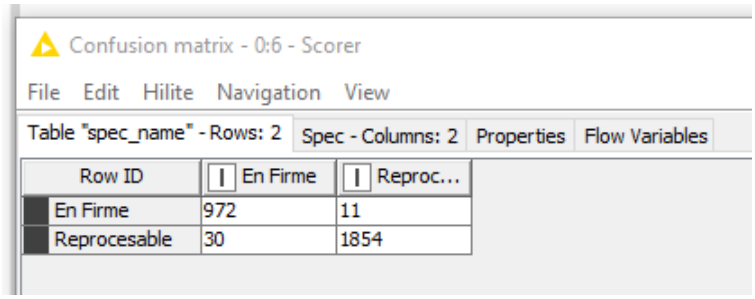

Figura 5.5. Resultados del modelo predictivo de este proyecto

Hay que destacar que, en el caso del desarrollo del modelo elegido tras la comparación, se han introducido una muestra completamente distinta a la introducida para realizar la comparación de modelos. Esta nueva muestra tiene un total de 2868 rechazos de entrada extraídos también del sistema operativo de la empresa, como se explica en el punto 4.1.5.

Como se puede observar, de 983 rechazos que se habían resuelto en firme, el modelo ha acertado en 972, fallando solo en la predicción de 11 rechazos. Del mismo modo, para los 1884 que se habían resuelto como reprocesables, el modelo acierta en 1854, fallando tan solo en 30. Ambos son índices muy buenos de acierto.

Además, de los 2868 rechazos que se le introdujeron como datos de entrada, ha predicho 2867 (983 en firme + 1884 reprocesables), es decir, todos menos uno, lo que lleva a pensar que una vez más se demuestra la calidad de predicción que ofrece este modelo.

En la Tabla 5.2 se muestra un extracto de la tabla completa que devuelve este Random Forest, a través del nodo Excel Writer.

Es un extracto porque solo se revelan algunos casos, debido a la gran dimensión de esta, además de mostrar solo los campos de “Rechazo:Rechazo”, que supone la identificación de rechazo, “Tipo Rechazo”, que corresponde a la solución que le dio el equipo realmente, “Prediction ()”, que corresponde a la solución que le ha dado el modelo, y “Prediction () (Confidence)”, que corresponde a la confianza con la que el modelo ofrece la medida, es decir la probabilidad que ha calculado de acierto en la decisión que ha tomado para cada caso.

Tabla 5.2. Fracción de resultados de predicción del modelo

Rechazo: Rechazo	Tipo Rechazo	Prediction ()	Prediction () (Confidence)
390172	En Firme	Reprocesable	59%
389952	Reprocesable	Reprocesable	100%
391003	Reprocesable	Reprocesable	99%
389027	En Firme	En Firme	100%
390117	Reprocesable	Reprocesable	100%
390077	Reprocesable	Reprocesable	99%
390078	Reprocesable	Reprocesable	92%
389690	Reprocesable	Reprocesable	85%
389838	Reprocesable	Reprocesable	100%
390075	Reprocesable	Reprocesable	100%
389650	Reprocesable	Reprocesable	100%
390086	Reprocesable	Reprocesable	100%
390150	Reprocesable	Reprocesable	93%
391889	Reprocesable	Reprocesable	100%
390546	Reprocesable	Reprocesable	98%
391043	En Firme	En Firme	98%
389989	Reprocesable	Reprocesable	98%
390047	Reprocesable	Reprocesable	100%
390056	Reprocesable	Reprocesable	99%
390553	Reprocesable	Reprocesable	94%
389981	Reprocesable	Reprocesable	99%
390192	Reprocesable	Reprocesable	69%
390206	En Firme	En Firme	79%
388479	Reprocesable	Reprocesable	100%
388478	Reprocesable	Reprocesable	100%
388101	En Firme	En Firme	100%
388148	Reprocesable	Reprocesable	93%
386576	En Firme	En Firme	79%
386625	Reprocesable	Reprocesable	100%
388175	Reprocesable	Reprocesable	94%
388178	En Firme	En Firme	100%
388177	En Firme	En Firme	91%
388859	En Firme	En Firme	100%
391991	Reprocesable	Reprocesable	100%
392016	Reprocesable	Reprocesable	100%
392015	Reprocesable	Reprocesable	100%
391999	Reprocesable	Reprocesable	100%
392003	Reprocesable	Reprocesable	100%
392000	Reprocesable	Reprocesable	100%
391992	Reprocesable	Reprocesable	100%
389287	En Firme	En Firme	100%
390249	Reprocesable	Reprocesable	76%
…	…	…	…

5.3 Presupuesto y análisis financiero.

En este capítulo se detallan los distintos apartados en los que se desglosan los costes del proyecto, agrupados en tres grandes bloques: recursos humanos, software, y recursos y accesorios.

Este desglose viene representado en la Tabla 5.3:

Tabla 5.3. Desglose de costes totales.

TOTAL			
RECURSOS HUMANOS			
Descripción	**Nº horas**	**€/hora**	**€**
Desarrollador	304	10,00	3040,00
Supervisor	12	15,00	180,00
		Total:	3220,00
SOFTWARE			
Producto	**Unidades**	**€/unidad**	**€**
KNIME - Licencia educativa	1	0,00	0,00
Windows 10 Pro	1	259,00	259,00
Microsoft Office 365	1	69,00	69,00
		Total:	328,00
EQUIPOS Y ACCESORIOS			
Descripción	**Unidades**	**€/unidad**	**€**
Ordenador portátil Lenovo	1	800,10	800,10
Monitor 23 pulgadas Lenovo	1	129,99	129,99
Cable HDMI 1 metro	1	6,99	6,99
Ratón inalámbrico Bluetooth	1	15,29	15,29
Teclado USB Logitech	1	20,99	20,99
		Total:	973,36
Total general:			**4521,36**

5.4 Tabla de ahorros.

Uno de los objetivos del modelo predictivo era conseguir ahorrar costes de operativa gracias a la automatización de los rechazos que se encuentren por encima del umbral determinado por el usuario.

De este modo, se ha confeccionado la Tabla 5.4, en la que se reflejan los porcentajes de ahorro con respecto al coste de la actividad de operación en función del umbral que se establezca como límite para la automatización de rechazos. Es decir, con los porcentajes de ahorro que ofrece la tabla, conociendo el coste que supone la actividad de resolución de rechazos, y conociendo el umbral que se ha establecido a partir del cual los rechazos se resolverán automáticamente en firme, se puede conocer el ahorro en coste que supone ducha automatización.

Hay que destacar que se ha elegido este formato de expresión del ahorro para no revelar los costes reales de la compañía por motivos de confidencialidad.

Tabla 5.4. Porcentajes de ahorro en base al umbral seleccionado.

UMBRAL MODELO	% Ahorro actividad
90%	0,017%
80%	0,978%
67%	4,335%
62%	5,021%
59%	5,054%
57%	5,341%
55%	5,412%
52%	5,453%
51%	5,474%
44%	6,548%
41%	6,652%
35%	6,666%
29%	6,722%
14%	12,802%
12%	13,356%
8%	13,501%
2%	13,591%

CAPÍTULO 6 - CONCLUSIONES Y TRABAJOS FUTUROS

6.1 Conclusiones.

En general, se puede concluir que el proyecto realizado ha resultado satisfactorio, debido a que se ha conseguido elaborar un modelo predictivo que ofrezca, con un elevado índice de fiabilidad, el resultado de si un rechazo acabará como en firme o como reprocesable.

Tanto el objetivo principal como los secundarios se han cumplido: con el punto de partida en el análisis del modelo de negocio del sector energético, y centrando el trabajo en conseguir un modelo predictivo de calidad, se han extraído y depurado los datos con especial interés y cuidado, consiguiendo una serie de instancias con características muy favorables para realizar con ellas una buena predicción. Además, se ha realizado una comparación de varios tipos de modelos llegando a la conclusión de que el Random Forest es el que mejor se adapta al problema en cuestión, justificando además así la elección tomada por parte de la empresa para lanzar su modelo. Y, en consecuencia, se ha obtenido un modelo ajustado y veraz, que cumple perfectamente los requisitos que se le pedían.

En definitiva, se han alcanzado todas las metas impuestas desde el principio del proyecto. El camino no ha sido sencillo, pero gracias al esfuerzo y dedicación empleados se han ido superando todas las barreras que se presentaban hasta llegar al final del recorrido.

Es importante destacar que no siempre es así, pues la incertidumbre es la reina en cualquier trabajo técnico. Siempre se sabe cómo se empieza y cuál es el objetivo, pero nunca se sabe cómo se va a acabar.

Por último, destacar y agradecer la labor de la empresa, pues todo ha sido posible gracias a su ayuda. Los medios, los sistemas, los datos, e incluso las personas que se han brindado a colaborar en el proyecto, son en gran medida también los responsables de que haya resultado exitoso.

6.2 Trabajos futuros.

Aunque con la implantación del modelo parezca que se ha llegado a la meta, es cierto que aún queda mucho camino por recorrer.

Empezando por el tipo de rechazos que cubre el modelo. Como se ha mencionado a lo largo del trabajo, este solo cubre los casos que se encuentran dentro del cost to serve, es decir, las altas de Energía XXI, y la modificaciones y bajas tanto de Energía XXI como de Endesa Energía. Dicho esto, aún quedan por introducir en el modelo los correspondientes al cost to adquiere, es decir, las altas de Endesa Energía. Este, por ejemplo, puede constituir el primero de los próximos avances a conseguir en el futuro.

En el segundo punto podría encontrarse el hecho de un automatismo total. En estos momentos, el modelo predictivo se ejecuta por un lado y el sistema de gestión de los rechazos por otro. De este modo, un operador tiene que extraer los datos del sistema de rechazos, cargarlos en el modelo, obtener las predicciones en base al umbral establecido, volver a cargar los casos que se encuentran dentro del umbral en el

sistema, y entonces este los finaliza en firme de forma automática. Realmente aún existe una gran parte de trabajo manual en el proceso. Lo realmente innovador sería el poder integrar el modelo predictivo en el propio sistema de gestión de rechazos. De ese modo, lo único que tendría que hacer el personal del equipo sería establecer el umbral por encima del cual los rechazos se resolverán en firme. Automáticamente, el modelo iría al apartado de rechazos pendiente de gestión, extraería los que cumplen los requisitos para poder predecirlos, ofrecería un resultado de probabilidad para cada uno de estos rechazos y resolvería en firme todos aquellos que se encuentren por encima del umbral. Así, la gestión manual sería mínima.

Y en el tercer punto, aunque sería muy desafiante, se encuentra el hecho de predecir el propio rechazo, antes incluso de lanzar la solicitud de la distribuidora. En cuanto el canal solicitara la operación y entrara en los servidores de la comercializadora, el modelo analizaría las solicitudes ofreciendo una probabilidad de que esta sea rechaza. De ese modo se podría establecer de nuevo un umbral a partir del cual, las solicitudes no sean trasladadas a la distribuidora si no que, debido a la alta probabilidad de que se convierta en un rechazo pase directamente al equipo para que lo valore y ajuste los parámetros necesarios antes de mandárselo a la distribuidora. El proceso es casi el mismo, solo que, en los casos de solicitudes con alta probabilidad de ser rechazadas, se ahorraría el tiempo que pasa entre que se envía a distribuidora, esta la analiza, la rechaza y la envía de vuelta. Con este avance, se ajustaría la solicitud antes de mandarla, reduciendo en gran medida la posibilidad del rechazo y agilizando mucho el flujo de trabajo.

En general, estas son unas ideas ambiciosas pero que están en el camino de lo que se busca, avanzar en automatismos que permitan realizar un trabajo más económico y sobre todo más eficiente.

CAPÍTULO 7 - BIBLIOGRAFÍA

1. Endesa. (2022). *Quienes somos*. Obtenido de https://www.endesa.com/es/sobre-endesa/quienes-somos
2. Bellot, A. (8 de Octubre de 2018). *Alcanzia. ¿Cuál es la distribuidora eléctrica de mi zona?* Obtenido de https://alcanzia.es/blog/distribuidoras-electricas-en-espana-por-zonas/
3. Patiño, M. Á. (26 de Enero de 2019). Así es la jungla de las 273 eléctricas que hay en España. *Expansión*.
4. Rosenblatt, F. (1957). The Perceptron, A Perceiving and Recognizing Automaton. *Cornell Aeronautical Laboratory*.
5. Rosenblatt, F. (1959). Two Theorems of Statistical Sepa-rability in the Perceptron. Symposium of the Mechanisation of Thought Processes. *National Physical Laboratory, Teddington*.
6. Rosenblatt, F. (1960). Perceptron Simulation Experiments. *Proc. Inst. Radio Engineers. V18*, 301-309.
7. Mosteller, R. B. (1951). A mathematical model for simple learning. *Psychological review.*, 313-323.
8. Widrow, B. (1961). Self-adaptive discrete systems, Theory Self Adapf. Contr. Syst. *IFAC Symp*.
9. Braverman, E. M. (1962). The experiments with training a machine to recognize patterns.
10. Bongard, M. M. (1961). Simulation of the recognition process on a digital computing machine. *Biophysics, vol.4*.
11. Tsypkin, Y. Z. (1966). Adaptation, Training And Self-Organization In Automatic Systems. *Automation And Remote Control, V.27*, 16-51.
12. Feldbaum, A. A. (1960). Dual Control Theory . *Automation and Remote Control, Vol.21*, 1960.
13. Papert, M. M. (1969). Perceptrons: An Introduction to Computational Geometry . *MIT Press*.
14. Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 193-202.
15. Rumelhart D. E., H. G. (1986). Learning representations by back-propagating errors. *Nature*, 533-536.
16. Dreyfus. (1990). Artificial Neural Networks, Back Propagation, and the Kelley-Bryson Gradient Procedure. *Journal of Guidance, Control, and Dynamics*, 926-928.
17. Woerner, P. W. (2017). Is Your Company Ready for a Digital Future? *MIT Sloan Management Review* .

18. Arana, R. (15 de Febrero de 2021). *TTANDEM DIGITAL STUDIO*. Obtenido de https://www.ttandem.com/blog/que-es-la-transformacion-digital-y-por-que-es-necesaria-para-cualquier-negocio/#:~:text=La%20transformaci%C3%B3n%20digital%20se%20puede,valor%20a%C3%B1adido%20a%20sus%20clientes.

19. Data, P. (Consultado el 02-06-2022). *POWER DATA*. Obtenido de https://www.powerdata.es/transformacion-digital

20. Fernández, A. (2022). Introducción a la ciencia de datos.

21. Julian Luengo, D. G.-G.-G. (2020). *Big Data Preprocessing. Enabling Smart Data.* Springer.

22. Fernández, A. (2022). Fundamentos de Machine Learning.

23. Han, J. K. (2011). *Data Mining: Concepts and techniques.* San Francisco, CA, USA: Morgan Kaufmann Publishers.

24. Fernández, A. (2022). Aprendizaje supervisad: Clasificación.

25. Witten, I. H. (2017). *Data mining: practical machine learning tools and techniques.* Amsterdam: Morgan Kaufmann.

26. KNIME. (s.f.). *Open for innovation, KNIME*. Obtenido de https://www.knime.com/

Resumen

Inserción de modelos predictivos en empresas del sector energético con motivo de la digitalización.

El presente documento recoge el desarrollo completo sobre la realización y aplicación de un modelo predictivo, que se encuadra dentro del ejercicio empresarial de las compañías dedicadas al sector energético, en este caso Endesa.
Con el objetivo principal de ayudar a la automatización y gestión eficiente de procesos productivos, se han definido los conceptos teóricos que fundamentan el caso, se ha realizado una organización progresiva de las tareas a llevar a cabo, y se ha terminado por el desarrollo práctico del modelo, teniendo siempre como referente uno ya existente desarrollado e implantado en la compañía. Esto ha permitido comparar los resultados para contrastar así la calidad de los mismos.
En general, se han conseguido alcanzar los objetivos propuestos, logrando poner en marcha un modelo predictivo que ofrece resultados muy positivos, aportando así un punto de valor añadido a la empresa y mejorando la calidad y eficiencia de la actividad de la misma.

Abstract

Insertion of predictive models in companies in the energy sector due to digitization.

This document includes the complete development on the realization and application of a predictive model, which falls within the business exercise of companies dedicated to the energy sector, in this case Endesa.
With the main objective of helping to automate and efficiently manage production processes, the theoretical concepts underlying the case have been defined, a progressive organization of the tasks to be carried out has been carried out, and the practical development of the model, always having an existing one developed and implemented in the company as a reference. This has made it possible to compare the results in order to contrast their quality.
In general, the proposed objectives have been achieved, managing to launch a predictive model that offers very positive results, thus providing a point of added value to the company and improving the quality and efficiency of its activity.