TÍTULO

ANÁLISIS DE DATOS Y DESARROLLO DE MODELO PREDICTIVO EN EL SECTOR ENERGÉTICO

AUTOR

Daniel González Rey

Esta edición electrónica ha sido realizada en 2024

Director	Dr. D. Pedro Javier Zarco Periñán
Instituciones	Universidad Internacional de Andalucía ; Universidad de Granada ; Universidad de Málaga ; Universidad de Almería
Curso	*Máster Universitario en Transformación Digital de Empresas (2022/23)*
©	Daniel González Rey
©	De esta edición: Universidad Internacional de Andalucía
Fecha documento	2023

TRABAJO FIN DE MÁSTER

Máster en Transformación digital de empresas

Universidad Internacional de Andalucía

"Análisis de datos y desarrollo de modelo predictivo en el sector energético"

Itinerario: Itinerario

Curso: 2022/2023

Modalidad: Técnico

Alumno/a: Daniel González Rey

Director/es:

- Pedro Javier Zarco Periñán

AGRADECIMIENTOS

Me gustaría aprovechar este capítulo de agradecimientos para mencionar a aquellas personas sin las cuales no podría haber llegado a donde estoy hoy. En primer lugar, a mi familia, mis padres me han apoyado y han dejado todo tanto personalmente como profesionalmente, mis abuelos con tanto cariño y apoyo incondicional, y amigos, muchas veces sacándome una sonrisa cuando lo necesitaba. Profesionalmente quiero agradecer a todos los profesores que han aportado su granito de arena para mi formación, especialmente aquellos que les apasionaba lo que hacían y dejaban parte de esa pasión en nosotros. Muchas gracias a Pedro, mi tutor de TFM, ha dedicado muchas horas de esfuerzo y trabajo para que este TFM pueda ser el que es, y sin él no sería el mismo. Finalmente quería dar las gracias a todo el equipo de Endesa, Luis, Marta, Raquel, Víctor, y tantos otros. El equipo de Endesa se ha volcado en mí desde que llegué y gracias a su trabajo, hoy puedo estar presentando este Trabajo Fin de Máster.

SIGLAS Y ACRÓNIMOS

Tabla de siglas y acrónimos.

Abreviatura	Concepto	Significado
IA	Inteligencia Artificial	
TFM	Trabajo Fin de Máster	
CSV	Coma Separate Values	Valores separados por comas
CNAE	Clasificación Nacional de Actividades Económicas	Código que sirve para identificar el tipo de negocio que tiene una empresa o un trabajador autónomo
CIE	Certificado de Instalación Eléctrica	Documento oficial que certifica que una instalación cumple con todos los requisitos necesarios para suministrar energía
	Modelo	Se refiere a aquella inteligencia artificial entrenada para resolver un problema
	Minimodelos	Hace referencia a aquellos modelos predictivos que usan pocas columnas para la predicción y son utilizados para predecir los motivos de rechazo de una solicitud
	Hosting	Servidor en la nube para alojar una aplicación web
	Framework	Marco o esquema de trabajo generalmente utilizado por programadores para realizar el desarrollo de software
	Solicitud	Conjunto de datos que son enviados para solicitar algo
	Pandas	Librería de código abierto de Python especializada en el manejo y análisis de estructuras de datos
	Scikit-learn	Biblioteca para aprendizaje automático de software libre para

		Python
	NumPy	Biblioteca de Python que se utiliza para trabajar con matrices, álgebra lineal, etc..
	TensorFlow	Biblioteca de código abierto de Python para facilitar la creación de redes neuronales y el aprendizaje profundo
	Tester	Aquella persona encargada de probar un producto software en busca de errores o fallos.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El presente documento detalla el desarrollo y la implementación de un modelo predictivo aplicando técnicas de Machine Learning y ciencia de datos, así como una aplicación web para su uso. Este modelo está destinado a predecir el resultado de solicitudes dentro del departamento de gestión de rechazos de Endesa, empresa encuadrada en el sector energético.

Durante el desarrollo de este trabajo se han definido una serie de bases teóricas y técnicas, necesarias para dicho trabajo. Además se ha detallado exhaustivamente la manera de implementar dicho modelo, así como la integración de la aplicación web. Finalmente se ha compartido un análisis de los resultados obtenidos y una conclusión acerca de los mismos y del trabajo en general.

En conclusión, se han conseguido alcanzar los objetivos propuestos, se ha logrado obtener un modelo predictivo de calidad integrado en una aplicación web útil y cómoda para el uso de los empleados. De esta manera se ha aportado un salto en la transformación digital de la empresa y en la automatización de procesos, que mejoran la calidad y eficiencia de los servicios.

Palabras clave: Modelo predictivo, machine learning, ciencia de datos, aplicación web, automatización, transformación digital, sector energético.

ABSTRACT

This document provides a detailed account of the development and implementation of a predictive model using Machine Learning and data science techniques, along with a web application for its utilization. This model is designed to predict the outcome of requests within the rejection management department of Endesa, a company operating in the energy sector.

Throughout the course of this work, a set of theoretical and technical foundations necessary for the project have been defined. Additionally, the implementation of the model and the integration of the web application have been thoroughly described. Finally, an analysis of the obtained results has been presented, along with a conclusion regarding these outcomes and the overall work.

In conclusion, the proposed objectives have been successfully achieved. A high-quality predictive model integrated into an easy-to-use web application for employee convenience has been developed. By doing so, a significant contribution to the company's digital transformation and process automation has been achieved, improving the quality and efficiency of the services provided.

Key words: Predictive model, machine learning, data science, web application, automation, digital transformation, energy sector.

ÍNDICE

CAPÍTULO 1 - INTRODUCCIÓN.

1.1 Motivación del trabajo fin de grado

Este Trabajo de Fin de Máster tiene varias motivaciones, la primera sería continuar la formación acerca de tratamiento de datos e inteligencia artificial. Este trabajo es muy fructífero para seguir investigando y aprendiendo acerca de técnicas de ingeniería de características y de machine learning que fueron introducidas previamente en el máster.

Otra motivación es la posibilidad de realizar un trabajo que pueda impactar realmente en una transformación digital para la empresa. Este trabajo permite observar cómo se lleva a cabo este proceso en una empresa real y cómo gestionar las dificultades imprevistas que puedan ir surgiendo.

Por último, este trabajo da la oportunidad de aportar realmente valor a la empresa dotándola de una tecnología para cubrir una necesidad real de la misma.

1.2 Objetivos

Los objetivos de este capítulo de Trabajo de Fin de Máster se centran en la implementación de un modelo predictivo utilizando el algoritmo de Random Forest con Django para abordar la gestión de rechazos en el sector de distribución de Endesa dentro del sector energético.

Concretamente, se tratará de que antes de que el departamento de gestión de rechazos envíe una solicitud a la empresa de distribución poder conocer si va a ser rechazada o no, y los posibles motivos de ello, de esta manera poder corregir estos antes de enviar dicha solicitud.

A continuación, se detallan más específicamente los objetivos a alcanzar:

- **Analizar el contexto de la gestión de rechazos en el sector de distribución de Endesa:** El primer objetivo es comprender en profundidad el marco y los desafíos asociados con la gestión de rechazos en el sector de distribución de energía eléctrica. Se realizará una revisión exhaustiva de la literatura relevante y se analizarán los procesos actuales de gestión de rechazos en Endesa, identificando las áreas de mejora y las oportunidades de implementación de un modelo predictivo.
- **Recopilar y preparar los datos necesarios:** Para implementar un modelo predictivo eficaz, es crucial contar con un conjunto de datos adecuado y bien preparado. En este objetivo, se realizará la recopilación de los datos históricos de rechazos en la distribución eléctrica de Endesa, considerando variables relevantes como la ubicación geográfica, el tipo de rechazo, la duración y el impacto en el suministro.

Además, se llevará a cabo la limpieza y transformación de los datos para garantizar su calidad y coherencia.

- **Desarrollar y entrenar el modelo predictivo utilizando el algoritmo de Random Forest:** Este objetivo implica la implementación del modelo predictivo utilizando el algoritmo de Random Forest. Se tratará de implementar un modelo predictivo para predecir los posibles rechazos que sufran las solicitudes, así como los motivos de rechazo de los mismos.
- **Evaluar y optimizar el rendimiento del modelo:** Una vez desarrollado el modelo predictivo, es necesario evaluar su rendimiento y realizar ajustes para lograr la mayor precisión posible. Se utilizarán técnicas de validación cruzada y métricas de evaluación adecuadas para medir la eficacia del modelo. En base a los resultados obtenidos, se realizarán optimizaciones, como la selección de características relevantes y la calibración de parámetros, para mejorar la calidad de las predicciones.
- **Integrar el modelo predictivo en una aplicación web utilizando Django:** Para facilitar su implementación y uso práctico, se desarrollará una aplicación web utilizando el framework Django. Esta aplicación permitirá a los usuarios de Endesa acceder y utilizar el modelo predictivo de manera intuitiva. Se diseñará una interfaz de usuario amigable que proporcione información sobre los rechazos esperados y las recomendaciones para su gestión, contribuyendo así a mejorar la eficiencia y la toma de decisiones en la gestión de rechazos.

1.3 Contexto

Para poner en contexto este Trabajo de Fin de Máster se indicará cómo es la empresa al que va destinado el sistema, cómo funciona el sector al que va aplicado y cómo es concretamente el día a día de la gestión de los rechazos, que es el entorno donde se aplicará esta tecnología.

1.3.1 Empresa destino.

Endesa es una empresa española dedicada al sector energético. Es una de las compañías líderes en la generación, distribución y comercialización de electricidad en España y en otros países. Endesa opera en toda la cadena de valor eléctrica, desde la generación de energía hasta su distribución y venta.

Endesa es la empresa líder del sector eléctrico español y el segundo operador del mercado eléctrico en Portugal. Cuenta con cerca de 10 mil empleados y presta servicio a más de 10 millones de clientes. Desarrolla su actividad fundamentalmente en el mercado de España y Portugal y en menor medida, comercializa electricidad y gas en otros mercados europeos.

Fundada en 1944, Endesa ha crecido y diversificado sus actividades a lo largo de los años. La empresa se dedica principalmente a la generación de energía eléctrica a través de una amplia gama de fuentes, incluyendo centrales térmicas, hidroeléctricas, eólicas y solares. Endesa también participa activamente en la distribución y comercialización de electricidad, llevando la energía a hogares, empresas e industrias.

La actividad de generación se realiza en España, Portugal y Marruecos. Endesa participa en centrales de producción eléctrica que funcionan a partir de diferentes fuentes de energía: hidroeléctrica, nuclear, térmica, eólica y solar. Además contribuye a la transición energética hacia una economía descarbonizada a través de la apuesta por las energías renovables, la digitalización y la economía circular.

1.3.2 Funcionamiento del sector.

Endesa se encuentra dentro del sector energético, un sector fundamental en el funcionamiento de cualquier sociedad desarrollada. El sector energético es el conjunto de actividades involucradas en la generación, distribución y comercialización de la energía. A continuación se indicará cómo funciona este sector en España.

La generación de energía en el sector energético de España cuenta con un sistema diversificado de generación que abarca diferentes fuentes de energía, tanto convencionales como renovables. En términos de generación convencional, España tiene una capacidad significativa de generación térmica, especialmente en plantas de ciclo combinado que utilizan gas natural como combustible. Respecto a energía nuclear, actualmente España cuenta con siete reactores nucleares en operación, esta energía ha sido fundamental para garantizar la seguridad del suministro eléctrico en el país. Además, España cuenta con una importante capacidad de generación hidroeléctrica aprovechando los recursos hídricos del país. En los últimos años, España ha experimentado un notable crecimiento en la generación de energía renovable. La energía eólica y la energía solar fotovoltaica han experimentado un desarrollo significativo, convirtiendo a España en uno de los líderes en la generación de energía renovable en Europa. El país cuenta con una amplia capacidad instalada de parques eólicos y plantas solares, aprovechando su clima favorable y su ubicación geográfica.

La distribución en el sector energético implica el transporte y suministro de energía desde las instalaciones de generación hasta los consumidores finales. En España, la distribución de energía se lleva a cabo a través de una extensa red de infraestructuras y sistemas de transporte. Esta red incluye líneas de transmisión de alta tensión y subestaciones que permiten la transferencia de energía eléctrica a larga distancia, así como redes de distribución de media y baja tensión que llevan la energía a los usuarios finales. El sistema de distribución en España está gestionado por empresas distribuidoras, que son responsables de operar y mantener la infraestructura. Estas empresas se encargan de garantizar el suministro continuo de energía, la gestión de la carga, la solución de averías y la conexión de nuevos usuarios a la red. En el caso del sector del gas natural, la distribución se realiza a través de una extensa red de gasoductos y estaciones de distribución. La distribución en el

sector energético de España se rige por regulaciones y normativas específicas para garantizar la calidad del suministro, la seguridad de las instalaciones y la protección del medio ambiente. Las empresas distribuidoras están sujetas a supervisión y control por parte de las autoridades reguladoras, como la Comisión Nacional de los Mercados y la Competencia (CNMC) y la Dirección General de Política Energética y Minas.

La comercialización en el sector energético se refiere al proceso de venta y gestión de la distribución a los consumidores finales. En España, la comercialización de energía está regulada por normativas y leyes que garantizan la transparencia, la competencia y la protección de los derechos de los consumidores. Este proceso implica la participación de diferentes actores completamente diferenciados, como comercializadoras de energía, distribuidoras y productores. Las comercializadoras de energía son empresas que adquieren la energía de los generadores y la venden a los consumidores finales. Estas empresas pueden ser tanto comercializadoras tradicionales, que ofrecen tarifas reguladas por el Gobierno, como comercializadoras en el mercado libre, que ofrecen una variedad de opciones y planes tarifarios.

En España, el sector energético se divide en dos segmentos principales: el mercado regulado y el mercado libre. Estos dos mercados ofrecen diferentes opciones de suministro y tarifas para los consumidores. El mercado regulado está supervisado y regulado por el Gobierno y los precios de la energía son fijados por el Gobierno y se actualizan periódicamente. El acceso al mercado regulado está limitado a unas condiciones determinadas por las autoridades, aunque actualmente pueden acceder el 95% de los usuarios. Estas tarifas pueden variar según la hora del día, ya que se aplican diferentes precios en función de la demanda de energía en cada periodo horario. En el caso del gas natural, también existe una tarifa regulada que se establece de forma similar. Por otro lado, el mercado libre ofrece una mayor flexibilidad y el acceso al mismo es completamente libre para cualquier usuario. Las comercializadoras en el mercado libre pueden ofrecer diferentes tipos de tarifas, opciones de precios fijos o indexados, servicios adicionales y descuentos.

1.3.3 Gestión de rechazos.

Este trabajo se centra en el ámbito de la gestión de rechazos. La gestión de rechazos se encarga de gestionar aquellas solicitudes que vengan rechazadas por parte de la distribuidora. Endesa es una comercializadora y la gestión de rechazos es un departamento de la empresa que se encarga de gestionar las peticiones que tengan que hacer los usuarios finales con la distribuidora. Por ejemplo, si un usuario necesita subir la potencia de su suministro, la gestión de rechazos enviará esa petición a la distribuidora para realizarla, si ésta responde que no es posible, ahí entra nuestro departamento.

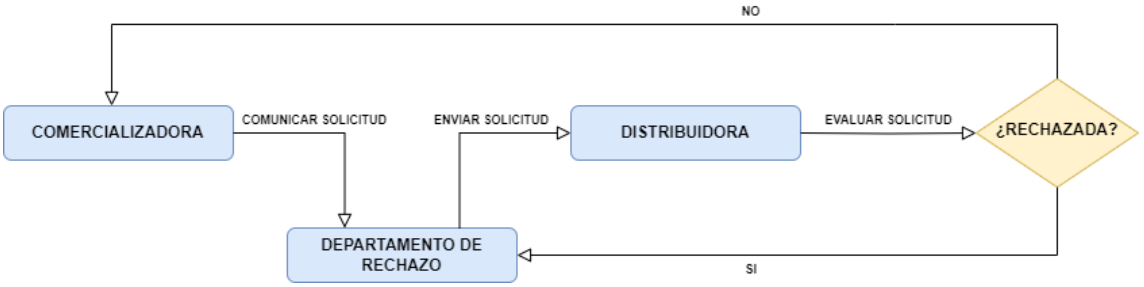
La gestión de rechazos cuando se recibe una solicitud que ha sido rechazada debe de revisar la solicitud enviada previamente y analizar cuál es el motivo del rechazo. Hay múltiples motivos de rechazos, desde un error administrativo, hasta un error técnico, o por problemas de contacto con el usuario final. Una vez que el departamento haya analizado el motivo por

el que ha sido rechazada, deberá de solucionarlo, si es posible, y volver a enviarlo a la distribuidora, o en caso de que no sea posible rechazarla en firme.

La gestión de rechazos tiene unos costes que quieren ser optimizados, es decir, minimizar el número de solicitudes que sean rechazadas y para aquellas que sean rechazadas tener que realizar las menores gestiones posibles. En esta optimización de procesos entra el presente trabajo, mediante un modelo predictivo se tratará de saber qué solicitudes serán rechazadas antes de ser enviadas, de manera que puedan corregirse previamente.

En la figura 1.1 se trata de representar el flujo que realiza una solicitud desde que es enviada de la comercializadora a la distribuidora, hasta que ésta es activada.

Figura 1.1. Flujo de una solicitud



1.4 Antecedentes.

En el ámbito de esta compañía, ya se ha implementado un modelo predictivo con el objetivo de optimizar el proceso de gestión de rechazos. En este caso el modelo se encarga de que una vez que han sido rechazados por parte de la distribuidora, determinar la probabilidad de ser rechazados en firme. En caso de que la probabilidad de ser rechazados sea de más del 80% el sistema determinará que es un rechazo en firme y así ahorrar el coste de gestión de esa solicitud.

En cuanto al tipo de modelo utilizado, se empleó un algoritmo de Random Forest. Aunque no es el modelo más complejo, se ha demostrado que tiene una alta precisión en las predicciones. La elección de este modelo se basa en la premisa de utilizar la opción más sencilla que cumpla con los requisitos de calidad de predicción.

Respecto al flujo de trabajo del modelo, este comienza con la extracción de los datos de entrada. El equipo encargado de los rechazos recopila los casos pendientes de gestión desde su sistema y estos datos se cargan en el modelo a través de una dirección web, utilizando un archivo CSV como medio de transferencia. Una vez cargados, los rechazos deben pasar por un filtro inicial para determinar si el tipo de rechazo ya está automatizado. Algunos motivos de rechazo se han establecido como procesos automatizados, por lo que el sistema verifica si algún rechazo pertenece a este grupo y lo excluye de la predicción, ya que se resolverá automáticamente. Para mejorar la precisión del modelo, se enriquece la información de entrada con datos adicionales que incluyen información sobre el producto contratado por el

cliente, la documentación asociada a la solicitud rechazada, contratos previos, ubicación del punto de suministro, entre otros. Una vez se ha enriquecido la información, el modelo predictivo entra en juego y proporciona los resultados deseados. Los umbrales de probabilidad se establecen en función de la decisión del encargado. En este caso,se decide un umbral del 80%, es decir, todos los rechazos con una probabilidad de ser en firme igual o mayor al 80% se gestionarán automáticamente, mientras que los que estén por debajo de dicho umbral serán revisados por los operadores.

1.5 Planificación

En la tabla 1.1 planificación temporal del trabajo, se detalla el tiempo que se ha destinado a cada una de las actividades y tareas necesarias para la realización de este Trabajo Fin de Máster.

Tabla 1.1. Planificación temporal del trabajo

Mes	Tarea o actividad	Duración (días)
Marzo	Extracción de conocimiento acerca de la empresa, el sector energético, la gestión de rechazos, etc..	3
	Conocimiento de las necesidades de la empresa	1
	Exploración de tecnologías para el desarrollo del trabajo	1
Abril	Obtención de datos para entrenamiento y predicción	1
	Preparación del entorno de desarrollo	1
	Formación acerca de técnicas de tratamiento de datos e ingeniería de características	2
	Comienzo de desarrollo de análisis y tratamiento de datos	1
Mayo	Desarrollo de análisis y tratamiento de datos	9
	Desarrollo de modelos predictivos	8
	Comienzo de redacción de la memoria	1
Junio	Desarrollo de aplicación web en Django	5
	Optimización de procesos	2
	Redacción de la memoria	4
Julio	Correcciones de la memoria	1

	Preparación de la defensa	2
		42

En total han sido necesarios 42 días, con una dedicación de 8 horas de media. Esto supone una dedicación de 336 horas.

1.6 Competencias utilizadas en el TFM

Durante la realización de este trabajo se han cubierto una serie de competencias. Estas competencias se clasifican entre básicas, generales, transversales y específicas. A continuación se detalla cómo se han cubierto cada una de ellas.

Respecto a las competencias básicas han sido cubiertas mediante la investigación acerca del Machine Learning para la realización de este proyecto, mediante la adquisición de nuevos conocimientos y técnicas. Estas competencias son las siguientes:

- 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 un 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.

Respecto a las competencias generales se han cubierto las mismas con la realización de este documento. En el que se han respetado todas las normativas y se ha realizado un trabajo exhaustivo de documentación. Las competencias generales son las siguientes :

- 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 científico-técnicos, 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.

Para cubrir las competencias transversales en este Trabajo Fin de Máster se ha seguido siempre una filosofía de aporte a la sociedad cooperando con todo tipo de profesionales. Estas competencias transversales cubiertas son las siguientes:

- 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.

Finalmente se han cubierto las competencias específicas, al realizar un producto software que sirve a la empresa para realizar análisis de datos, mejora en toma de decisiones y proporciona una mejora en el proceso productivo de la misma. Estas competencias son:

- 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.
- 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.
- CEI 1/2/3/4 Saber analizar y proponer nuevas soluciones tecnológicas y de mejora en el ámbito del Sector elegido de entre los cuatro itinerarios.

1.7 Estructura de la memoria del TFM

Este trabajo se ha estructurado en 7 capítulos, cada uno de ellos centrado en detallar un ámbito de los conocimientos y tareas necesarios para la realización de este Trabajo Fin de Máster. A continuación se detallan cada uno de ellos:

- **Capítulo 1 - Introducción:** Recoge las motivaciones, objetivos y planificaciones necesarias para el desarrollo del trabajo. Además se trata de poner en contexto el

trabajo y conocer sus antecedentes, para poder entender la extensión y sentido del mismo.

- **Capítulo 2 - Bases teóricas:** Se indican aquellos conceptos y conocimientos que han servido de base para la realización de este Trabajo Fin de Máster.
- **Capítulo 3 - Bases técnicas:** Se especifican aquellas tecnologías y demás herramientas necesarias para el desarrollo de este trabajo. Dichas tecnologías son los cimientos del modelo y la aplicación web desarrollados.
- **Capítulo 4 - Implementación:** Este capítulo trata de contar cómo se ha llevado a cabo el desarrollo del modelo predictivo, con su posterior integración en la aplicación web, así como mostrar el resultado visual de los mismos.
- **Capítulo 5 - Resultados y pruebas:** Se exponen los diferentes resultados que se han obtenido del modelo predictivo, así como un análisis de su eficiencia y configuración hasta decidir la opción óptima.
- **Capítulo 6 - Conclusiones y trabajos futuros:** Se presentan las conclusiones generales del proyecto y posibles caminos y proyectos que se puedan abordar a partir del mismo en el futuro.
- **Capítulo 7 - Bibliografía:** Recoge todos los documentos y fuentes bibliográficas de las que se ha servido este trabajo.

CAPÍTULO 2 - BASES TEÓRICAS

2.1 Transformación digital.

La transformación digital es un proceso integral que implica la aplicación estratégica de tecnologías digitales y cambios organizativos en una empresa u organización con el fin de optimizar sus operaciones, mejorar la experiencia del cliente y mantener su relevancia en un entorno empresarial cada vez más digitalizado. Se trata de una iniciativa que abarca todas las áreas de una organización, desde la gestión interna hasta la relación con los clientes y proveedores.

La transformación digital implica el aprovechamiento de las tecnologías digitales emergentes, como la inteligencia artificial, el análisis de datos, la computación en la nube, el Internet de las cosas y la realidad virtual, entre otras, para impulsar la eficiencia, la innovación y la competitividad. Estas tecnologías permiten a las organizaciones recopilar, analizar y utilizar grandes cantidades de datos en tiempo real, lo que facilita la toma de decisiones más informadas y basadas en evidencias.

Sin embargo, la transformación digital no se trata sólo de implementar tecnología, sino de promover un cambio cultural y organizativo en toda la empresa. Implica repensar los procesos y modelos de negocio existentes, fomentar la colaboración y la agilidad, y capacitar a los empleados para adquirir nuevas habilidades digitales. Además, la transformación digital también implica considerar las necesidades y expectativas cambiantes de los clientes y adaptar los productos, servicios y canales de distribución en consecuencia.

Este trabajo se centrará en la transformación digital por medio de la inteligencia artificial y el análisis de datos. Para ello se utilizarán los fundamentos de la ciencia de datos y se implementarán métodos como el machine learning.

2.2 Inteligencia artificial.

La inteligencia artificial (IA) es un campo de estudio y desarrollo que busca dotar a las máquinas y sistemas informáticos de habilidades y capacidades similares a las humanas, como el razonamiento, el aprendizaje, la percepción, la comprensión del lenguaje natural y la toma de decisiones. El objetivo de la IA es crear sistemas inteligentes capaces de realizar tareas de manera autónoma, adaptativa y eficiente. Concretamente, la IA puede incorporar técnicas de razonamiento lógico, planificación y procesamiento del lenguaje natural.

La IA se aplica en una amplia gama de industrias y sectores, como la medicina, la automoción, la manufactura, el comercio electrónico, la seguridad, la robótica y muchas otras. Sus aplicaciones van desde la detección de enfermedades y el diagnóstico médico hasta la conducción autónoma de vehículos, la personalización de recomendaciones en línea y la optimización de la cadena de suministro.

Existen diferentes enfoques y subcampos dentro de la IA, centrándose este trabajo en el aprendizaje automático (Machine Learning). El machine learning se enfoca en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender de los datos sin ser programadas explícitamente. La extracción de conocimiento de esos datos es fundamental por lo que será necesario conocer y aplicar los fundamentos de la ciencia de datos. Además, el aprendizaje automático incluye técnicas como las redes neuronales artificiales, que imitan el funcionamiento del cerebro humano, y los algoritmos de aprendizaje supervisado y no supervisado, que permiten a las máquinas reconocer patrones y realizar predicciones.

Otro subcampo de la IA es el procesamiento del lenguaje natural, que se ocupa de la comprensión y generación de texto y habla. Esto implica el desarrollo de sistemas capaces de entender y generar lenguaje humano. No obstante, este otro subcampo no será tocado en este trabajo.

Sin embargo, la IA también plantea desafíos y consideraciones éticas. Estos incluyen preocupaciones sobre la privacidad y la seguridad de los datos, la discriminación algorítmica, la falta de transparencia en la toma de decisiones de las máquinas y el impacto en el empleo y la sociedad en general.

2.3 Ciencia de datos e ingeniería de características.

La ciencia de datos es un campo multidisciplinario que combina elementos de matemáticas, estadística, informática y dominios específicos de conocimiento con el objetivo de extraer conocimientos y perspectivas útiles a partir de grandes volúmenes de datos. Se centra en el proceso de recopilación, almacenamiento, procesamiento, análisis y visualización de datos para descubrir patrones, tendencias y relaciones que puedan utilizarse para la toma de decisiones informadas y la resolución de problemas complejos.

En el contexto actual, donde se generan cantidades masivas de datos en diversas formas y formatos, la ciencia de datos desempeña un papel fundamental. Su objetivo es aprovechar el potencial de los datos para obtener información valiosa y acciones concretas. Esto implica aplicar técnicas y algoritmos avanzados para procesar y analizar datos, identificar correlaciones, realizar predicciones y construir modelos predictivos.

También es importante el concepto de ingeniería de características, este se enfoca en la creación y selección de variables o características relevantes y significativas a partir de los datos disponibles. Este proceso implica la transformación y manipulación de los datos brutos para extraer información relevante que pueda ser utilizada por los modelos predictivos. La ingeniería de características implica la aplicación de conocimientos de dominio, creatividad y experiencia para seleccionar, combinar o derivar nuevas características que puedan mejorar la calidad y el rendimiento de los modelos.

La ciencia de datos y la ingeniería de características son dos disciplinas estrechamente relacionadas que trabajan juntas para aprovechar al máximo los datos en un proyecto de análisis o modelado predictivo. La relación entre la ciencia de datos y la ingeniería de

características es bidireccional y complementaria. La ingeniería de características proporciona las entradas necesarias para el análisis de datos y la construcción de modelos, ya que los modelos solo pueden aprovechar la información que se les proporciona a través de las características. Al mismo tiempo, la ciencia de datos ayuda a guiar el proceso de ingeniería de características al evaluar la importancia y el impacto de diferentes características en la predicción y al proporcionar retroalimentación sobre la calidad y la relevancia de las características seleccionadas.

2.4 Machine Learning y modelos predictivos.

El aprendizaje automático, también conocido como machine learning, es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender de los datos y mejorar su rendimiento a través de la experiencia. En lugar de ser programados explícitamente para llevar a cabo tareas específicas, los sistemas de aprendizaje automático son capaces de aprender y adaptarse automáticamente a partir de los datos, identificando patrones y relaciones que les permiten realizar predicciones o tomar decisiones informadas.

El aprendizaje automático se basa en la idea de que las máquinas pueden analizar grandes volúmenes de datos y extraer información valiosa de ellos. Los algoritmos de aprendizaje automático pueden ser entrenados utilizando conjuntos de datos de entrenamiento, donde se les proporciona información de entrada y se les indica la salida deseada. A través del análisis de estos datos, los algoritmos aprenden a reconocer patrones y correlaciones entre las variables, lo que les permite realizar predicciones o clasificar nuevos datos en función de lo que han aprendido.

Por otro lado, los modelos predictivos son una aplicación clave del machine learning, ya que utilizan algoritmos y técnicas para predecir valores o eventos futuros basados en datos históricos. Este tipo de modelo se basa en la idea de que los patrones identificados en los datos pasados pueden ayudar a predecir tendencias o resultados futuros.

La relación entre el machine learning y los modelos predictivos es estrecha y fundamental. Los modelos predictivos son construidos utilizando algoritmos de machine learning, que se entrenan con grandes conjuntos de datos. Estos modelos pueden ser utilizados para hacer predicciones sobre nuevos datos o eventos que aún no han ocurrido.

El machine learning ofrece diferentes enfoques y algoritmos para construir modelos predictivos, como la regresión lineal, los árboles de decisión, las máquinas de vectores de soporte (SVM), las redes neuronales, entre otros. Cada algoritmo tiene sus propias características y aplicaciones, y la elección del algoritmo depende del problema específico y los datos disponibles.

En este caso, para el problema presentado se va a utilizar el algoritmo Random Forest, que es una extensión del algoritmo de árboles de decisión que combina múltiples árboles de decisión en un conjunto para mejorar la precisión y la generalización del modelo. En lugar de

basarse en un solo árbol de decisión, el algoritmo Random Forest crea un conjunto de árboles de decisión independientes, donde cada árbol se entrena con una muestra aleatoria de los datos de entrenamiento y utilizando un subconjunto aleatorio de las características disponibles. Esta aleatoriedad en la selección de datos y características reduce el riesgo de sobreajuste y mejora la capacidad de generalización del modelo.

Una vez que se ha seleccionado el algoritmo de machine learning adecuado, se realiza el proceso de entrenamiento del modelo. Durante el entrenamiento, el modelo se ajusta a los datos históricos para aprender las relaciones y patrones existentes, de manera que pueda hacer predicciones precisas en el futuro. Y una vez entrenado, el modelo predictivo se utiliza para realizar predicciones con datos nuevos o eventos futuros. El modelo aplica las relaciones aprendidas durante el entrenamiento para generar predicciones basadas en las características o variables de entrada proporcionadas.

CAPÍTULO 3 - BASES TÉCNICAS

3.1 Python

Para implementar el modelo predictivo se ha decidido el uso de Python como lenguaje principal. Python es un lenguaje de programación versátil y ampliamente utilizado en el ámbito del análisis de datos y el aprendizaje automático debido a su gran cantidad de bibliotecas especializadas y su facilidad de uso.

Python es especialmente adecuado para implementar modelos predictivos debido a su sintaxis clara y legible, lo que facilita el desarrollo y mantenimiento del código. Además, cuenta con una amplia gama de bibliotecas y frameworks, como NumPy, Pandas, Scikit-learn y TensorFlow, que proporcionan herramientas y funciones específicas para el procesamiento de datos y la construcción de modelos predictivos.

Una de las ventajas de utilizar Python para implementar un modelo predictivo es su capacidad para manejar grandes volúmenes de datos. Las bibliotecas mencionadas anteriormente ofrecen estructuras de datos eficientes y funciones optimizadas que aceleran el procesamiento y análisis de datos. Esto es especialmente útil cuando se trabaja con conjuntos de datos complejos o de gran escala.

Otra razón para utilizar Python es su fuerte integración con otras tecnologías y bibliotecas. Por ejemplo, es posible combinar Python con bibliotecas de visualización como Matplotlib o Seaborn para representar gráficamente los resultados del modelo y comunicar de manera efectiva los hallazgos. Además, Python se puede integrar con herramientas de desarrollo web y bases de datos, permitiendo la creación de aplicaciones en línea o sistemas en tiempo real basados en modelos predictivos.

La comunidad de Python es otra fortaleza importante. Existe una gran cantidad de recursos en línea, tutoriales, ejemplos de código y foros de discusión donde los desarrolladores pueden buscar ayuda y compartir conocimientos. Además, Python cuenta con una comunidad activa de científicos de datos y desarrolladores de aprendizaje automático, lo que fomenta la colaboración y la mejora continua de las bibliotecas y herramientas relacionadas.

3.2 Django

Se enfocó Django como marco de desarrollo web para implementar el modelo predictivo y la aplicación web necesaria. Django es un framework de Python ampliamente utilizado y de código abierto que permite desarrollar aplicaciones web de manera eficiente y robusta. Además, una de las razones principales para utilizar Django es su capacidad para manejar la lógica del servidor de manera organizada y escalable.

Django proporciona una arquitectura basada en el patrón de diseño Modelo-Vista-Controlador (MVC), para promover la separación entre las piezas de una aplicación y facilitar hacer cambios en un lugar particular de la aplicación sin afectar otras piezas. La M hace referencia a la porción de acceso a la base de datos, la V es la porción que selecciona qué datos mostrar y cómo mostrarlos, y la C la porción de navegación por parte del usuario.

Al implementar un modelo predictivo, Django se puede utilizar para construir una interfaz web que permita a los usuarios interactuar con el modelo y obtener predicciones en tiempo real. Django proporciona una estructura para definir modelos de datos, vistas y plantillas, lo que simplifica el desarrollo de la interfaz de usuario y la comunicación con el modelo predictivo.

Además, Django ofrece características de seguridad y autenticación integradas que garantizan la protección de los datos y el acceso controlado a la aplicación. Esto es especialmente importante en el contexto de un modelo predictivo, donde la privacidad y la seguridad de los datos son consideraciones críticas.

La gran comunidad activa de Django es otro factor a tener en cuenta. Existen numerosos recursos en línea, documentación detallada y una gran cantidad de paquetes adicionales disponibles, lo que facilita el desarrollo y la resolución de problemas en el proceso de implementación del modelo predictivo.

3.3 PythonAnywhere

Se utilizará PythonAnywhere como servicio de hosting para alojar la aplicación web que debe correr el modelo predictivo. PythonAnywhere es una plataforma en la nube que permite ejecutar aplicaciones y scripts de Python en línea, lo que lo convierte en una opción conveniente y eficiente para implementar un modelo predictivo en producción.

Una de las principales ventajas de utilizar PythonAnywhere es su facilidad de uso. Proporciona una interfaz de usuario intuitiva y amigable que simplifica el proceso de implementación y administración de aplicaciones web. Además, no se requiere configurar y mantener una infraestructura de servidor propia, lo que ahorra tiempo y recursos.

PythonAnywhere es compatible con la mayoría de los frameworks de Python, como Django o Flask, lo que permite desarrollar y ejecutar aplicaciones web de manera rápida y sencilla. Esto es especialmente útil al implementar un modelo predictivo, ya que se puede utilizar un framework para construir una interfaz de usuario y exponer el modelo a través de una API o página web.

La escalabilidad es otro aspecto a considerar al utilizar PythonAnywhere. El servicio se adapta a las necesidades del proyecto, permitiendo escalar vertical o horizontalmente según sea necesario. Esto es especialmente útil cuando se espera un aumento en la demanda o cuando se necesita procesar grandes volúmenes de datos en tiempo real.

PythonAnywhere también proporciona medidas de seguridad para proteger los datos y las aplicaciones. Utiliza protocolos de encriptación y autenticación seguros para garantizar la confidencialidad y la integridad de los datos transmitidos. Además, se realizan copias de seguridad regulares para garantizar la disponibilidad y la recuperación de datos en caso de fallos.

3.4 Microsoft Power BI

Microsoft Power BI es una plataforma de análisis de datos empresariales que permite visualizar, analizar y compartir información de manera efectiva. Se utilizará Microsoft Power BI para realizar el análisis de datos y la ingeniería de características. Las principales ventajas de utilizar Power BI para visualización de datos es su capacidad de visualización interactiva y dinámica. Power BI ofrece una amplia gama de visualizaciones gráficas y tabulares que permiten representar los resultados del modelo de manera intuitiva y comprensible. Estas visualizaciones pueden ser personalizadas y actualizadas en tiempo real, lo que facilita la exploración de los datos y la identificación de patrones o tendencias.

Power BI también proporciona herramientas de análisis avanzadas que permiten realizar cálculos, agregar datos y aplicar filtros para profundizar en los resultados del modelo predictivo. Estas capacidades analíticas nos permiten realizar un análisis detallado y obtener información adicional a partir de los resultados del modelo. Power BI también permite crear paneles de control interactivos y cuadros de mando personalizados. Estos paneles de control pueden incluir visualizaciones, informes y métricas clave relacionadas con el modelo predictivo. Esto es especialmente útil para los usuarios finales, ya que les brinda una vista rápida y comprensible de los resultados del modelo y les permite tomar decisiones informadas basadas en los datos.

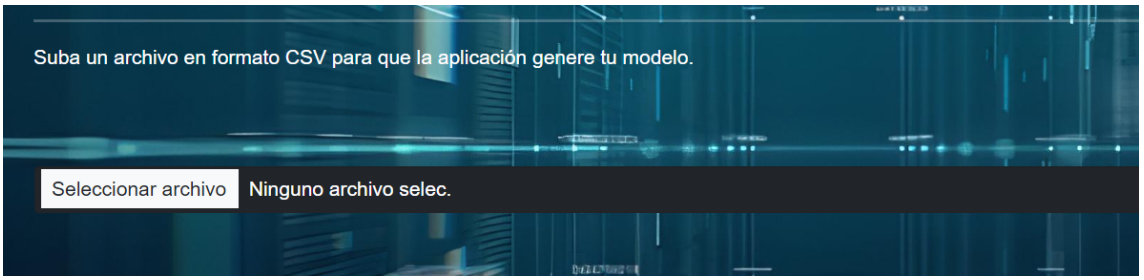
CAPÍTULO 4 - IMPLEMENTACIÓN

4.1 Obtención de los datos.

La obtención de datos es un paso crítico en el proceso de desarrollo de un modelo predictivo, ya que la calidad y la relevancia de los datos utilizados tendrán un impacto directo en la precisión y la efectividad del modelo.

Para obtener estos datos el usuario debe enviar un archivo en formato csv con la información con la que se va a entrenar el modelo.

Figura 4.1. Envío de fichero csv.



Mediante Python y la librería Pandas se recibe esa información y se trata conforme al formato correspondiente.

4.2 Limpieza de datos.

La limpieza de datos es esencial para garantizar la calidad y la integridad de los datos utilizados en el modelo. Los datos pueden contener errores introducidos durante la recopilación, el almacenamiento o el procesamiento, lo que puede afectar negativamente a la precisión y el rendimiento del modelo. Por lo tanto, es fundamental abordar estos problemas antes de utilizar los datos para el entrenamiento y la evaluación del modelo.

Más concretamente en este primer paso, se tratan los elementos nulos, los cuales son transformados en un 0 para poder ser tratados posteriormente por el modelo. Los elementos nulos no pueden ser tratados por un algoritmo como Random Forest. En la figura 4.2 se muestra a modo de ejemplo cómo se realiza.

Figura 4.2. Transformar vacíos.

```
def vaciosPor0(csv):
    for columna in csv.columns:
        csv[columna].fillna(0, inplace=True)

    return csv
```

Además, la columna a predecir, en este caso el estado de la solicitud, es un valor que puede tomar 'Activada' o 'Rechazada'. Para poder tratarla por el algoritmo como una columna válida, dicha columna es transformada de la siguiente manera, aquellas que son activadas se ponen a 0 y las rechazadas a 1.

Por último, al realizar un último análisis del fichero se observa que hay en los datos valores no válidos dentro de algunas columnas, valores como 'Desconocido' o guiones. Este tipo de valores no se deben tener en cuenta en el modelo como uno más, ya que lo que quieren representar realmente es que son un valor vacío o nulo. Por ello, ese tipo de valores se definen como 0 al igual que se realizó con los vacíos.

4.3 Ingeniería de características.

La ingeniería de características es un proceso fundamental en el desarrollo de un modelo predictivo, ya que tiene como objetivo seleccionar, crear y transformar variables o características de los datos que sean más informativas y relevantes para la tarea de predicción.

Como primer paso se seleccionan aquellas columnas que se consideran más interesantes y borrar todas aquellas que no vayan a ser utilizadas. Esta elección de columnas proviene de un análisis del negocio mediante reuniones con expertos en el sector, estudio de los valores de cada columna y las diferentes pruebas realizadas.

Además de la selección de columnas, se detecta que algunas columnas son categóricas y no ordinales, por lo que a priori no son válidas para el modelo. Para ello se deben realizar transformaciones de esas columnas y existiendo varios métodos para ello. A continuación se describen cuales pueden ser estos:

- **Codificación one-hot (binaria):** Este enfoque consiste en crear una nueva columna binaria para cada categoría en la columna original. Cada columna binaria representa la presencia o ausencia de esa categoría en una observación. En este caso en el modelo será utilizado para el campo "**tx_tlfn_contacto"** que representa números de teléfono, para saber si el usuario había rellenado o no ese campo.
- **Codificación ordinal:** En casos donde las categorías tienen un orden intrínseco, se puede asignar un valor numérico a cada categoría según su posición en la secuencia ordenada. Esto conserva la relación de orden entre las categorías y puede ser útil para algoritmos como el de árboles de decisión, ya que al tener que ramificar pueden hacerlo con más facilidad. En el conjunto de datos no fue encontrada ninguna columna que necesitara hacer uso de este método.
- **Codificación de frecuencia:** En este método, se asigna a cada categoría un valor numérico basado en la frecuencia de aparición en la columna. Las categorías más frecuentes obtienen valores más altos, mientras que las menos frecuentes obtienen valores más bajos. Esta codificación puede ayudar a capturar la importancia relativa

de cada categoría en el conjunto de datos. En el conjunto de datos se encontraron columnas para su utilización, que fueron los campos "**de_entlpobl_stro**" y "**de_prov_stro**" que representan el municipio y la provincia.

- **Codificación target:** Este método utiliza información sobre la variable objetivo para asignar valores numéricos a las categorías. Esto puede capturar la relación entre las categorías y la variable objetivo, pero también puede introducir riesgos de filtración de datos si no se realiza correctamente. Esto es lo que fue realizado en la limpieza de datos con la columna objetivo.
- **Codificación de impacto:** Este método utiliza una métrica de impacto, como el índice de información mutua o la ganancia de información, para asignar un valor numérico a cada categoría. La idea es medir cuánta información aporta cada categoría a la variable objetivo y asignar valores en función de su impacto. En este caso concretamente es utilizado el porcentaje de rechazos que tiene esa columna en función de cada categoría, por ejemplo, para la provincia de Cádiz el 25% es rechazada. Este método es muy útil para las columnas de distribuidora y del canal de recepción de solicitudes. Por lo que fue implementado en los campos "**de_distrib**" y "**de_canal**" para representar la distribuidora y el canal de comunicación.

Hay multitud de métodos para transformar columnas categóricas no ordinales y utilizarlas en modelos predictivos. La elección del método dependerá del conjunto de datos con el que se esté trabajando, la relación entre las categorías y la variable objetivo, y el algoritmo de aprendizaje automático que se utilizará.

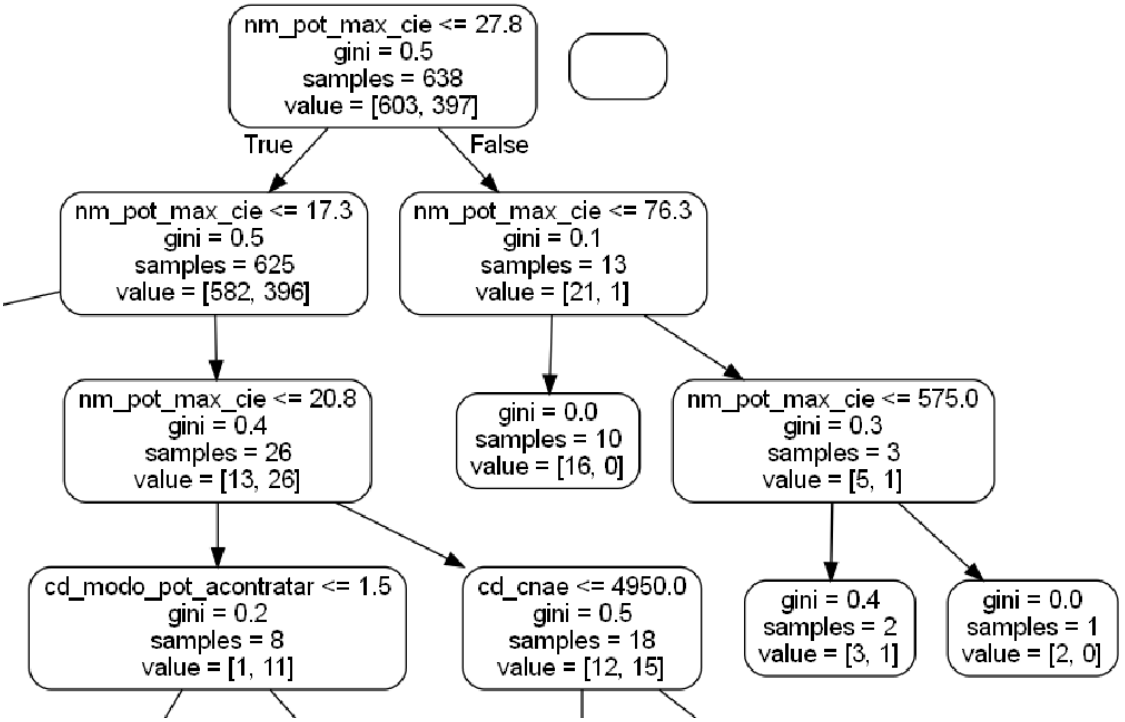
4.4 Implementación del Random Forest.

El algoritmo de Random Forest es ampliamente utilizado en el ámbito de la ciencia de datos y el aprendizaje automático debido a su capacidad para realizar predicciones precisas y lidiar con conjuntos de datos complejos. El Random Forest es un algoritmo de aprendizaje supervisado que combina múltiples árboles de decisión para formar un modelo predictivo robusto. En lugar de utilizar un solo árbol de decisión, el algoritmo de Random Forest construye múltiples árboles de decisión de forma independiente.

Cada árbol se construye eligiendo aleatoriamente una muestra de entrenamiento de tus datos, a partir de esa muestra de entrenamiento, se construye un árbol de decisión dividiendo los datos en función de las variables más importantes para clasificar o predecir. Durante la construcción del árbol, se realiza una selección aleatoria de características para evaluar en cada división del árbol. Esto implica que en cada nodo del árbol, solo se consideran un subconjunto aleatorio de características para tomar decisiones. Se construyen múltiples árboles de decisión, cada uno basado en diferentes muestras de entrenamiento y subconjuntos de características aleatorias. Para hacer predicciones, se utiliza el conjunto de árboles construidos. Cada árbol genera una predicción y, en el caso de la clasificación, se

toma la clase más frecuente entre los árboles. En la figura 4.3 se trata de mostrar un ejemplo real de una parte de un árbol de decisión para el conjunto de datos con el que se está trabajando.

Figura 4.3. Ejemplo de árbol de decisión real.



La principal ventaja del Random Forest radica en su capacidad para reducir el sesgo y la varianza inherentes a los árboles de decisión individuales. Al combinar múltiples árboles, el Random Forest puede promediar las predicciones de cada árbol, lo que resulta en una predicción más estable y precisa. Además, el algoritmo también realiza selecciones aleatorias de características durante la construcción de los árboles, lo que ayuda a reducir la dependencia de un subconjunto particular de características y evita el sobreajuste.

El Random Forest también proporciona información sobre la importancia relativa de las características en el proceso de predicción. Estas métricas permiten identificar las características más relevantes para el problema en cuestión y pueden ayudar en la interpretación del modelo.

En términos de implementación, el Random Forest está disponible en varias bibliotecas y herramientas populares de Python, como scikit-learn. Estas bibliotecas ofrecen una interfaz fácil de usar para construir, entrenar y evaluar modelos de Random Forest. Además, también brindan opciones para ajustar los parámetros del algoritmo, como el número de árboles y la profundidad máxima, lo que permite ajustar el modelo según las necesidades específicas del problema.

4.5 Motivos de rechazo

Además del modelo predictivo a realizar con el algoritmo Random Forest, se quiere ofrecer al usuario los motivos de rechazo por los cuales se predice que pueda estar rechazada esa solicitud. A continuación se detallarán cuáles son estos motivos, cómo se han definido y cómo funciona.

4.5.1 Definición de motivos

Para definir cuáles son los motivos que son importantes para el conjunto de datos a predecir, es importante el contacto con un experto en esos datos. Mediante un análisis expertos y diferentes pruebas de resultados se detallan los siguientes motivos:

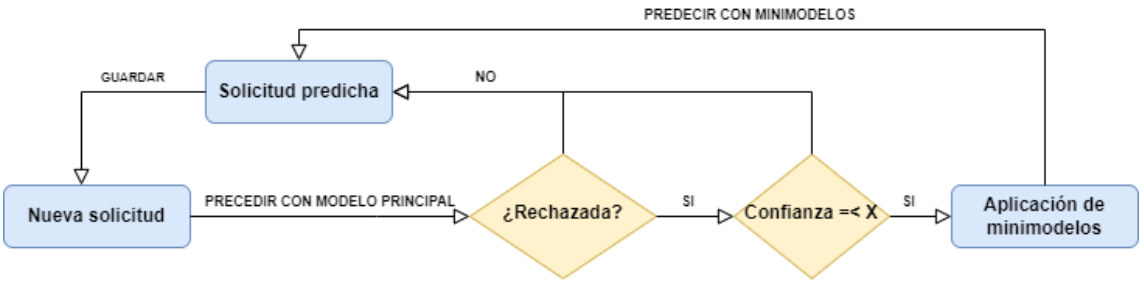
- **Documentación:** Hace referencia a los documentos asociados que tiene esa solicitud. Si se tiene una alta confianza de rechazo por ese motivo debe de producirse por falta de documentación técnica o administrativa. Para ello se realiza un modelo con dos columnas 'tiene_tecnico' y 'tiene_administrativo'.
- **Teléfono de contacto:** Todas las solicitudes portan información acerca del cliente que la solicita, en algunas de ellas aparece un número de teléfono para contactar con él. Al tener una alta confianza de rechazo por este motivo significa que la distribuidora puede que requiera este número para contactar con el cliente. El modelo para este motivo únicamente tiene la columna 'tx_tlfn_contacto'.
- **Potencia máxima por CIE:** Cuando se solicita una subida de potencia a una distribuidora hay cierta cantidad que no puede ser realizada por la propia instalación de la vivienda. Cuando se tenga una alta confianza de rechazo por este motivo querrá decir que esa restricción de potencia suele dar problemas y habrá que revisar la potencia solicitada por el cliente. Para el modelo se utiliza la columna 'nm_pot_max_cie'.
- **Distribuidora:** Las solicitudes son realizadas a multitud de diferentes distribuidoras diferentes hay algunas más conocidas y otras menos, y a veces las menos comunes pueden tener una manera especial de hacer las gestiones. Por ello si se tiene alta confianza de rechazo por la distribuidora, deben de fijarse en cuál es la misma, para ver si hay que hacer algunas gestiones especiales para ella y evitar el posible rechazo. Este modelo únicamente utiliza la columna 'de_distrib'.

4.5.2 Funcionamiento

El funcionamiento del sistema consiste en estimar los motivos por los que una solicitud puede venir rechazada, para estimar estos motivos de rechazo a esa solicitud se le aplicaran diferentes “minimodelos”. Estos "minimodelos" se tratan de modelos entrenados para cada uno de los motivos utilizando las columnas que son consideradas importantes para ese motivo concreto. Serán definidos tantos modelos como motivos haya, de esta manera al predecir una solicitud se realizará la predicción con cada uno de estos "minimodelos" y mediante la confianza de rechazo que nos muestren cada uno de ellos, se podrá saber cuales son los motivos de rechazo más probables para esa solicitud.

El proceso de predicción, por tanto, será el definido en la figura 4.4 proceso de predicción mediante modelado. Primero se realizará la predicción con el modelo principal obteniendo una predicción. Si la predicción es 'Activada', entonces se habrá terminado con esa solicitud, en caso de ser 'Rechazada' se continúa con el proceso. Si es 'Rechazada' y cumple un cierto porcentaje de confianza que se definirá en el capítulo de resultados y pruebas, entonces se aplican los "minimodelos". Aplicar los "minimodelos" implicará predecir esa solicitud con el modelo de cada motivo y devolver una lista con la confianza de rechazo para cada motivo. De esta manera el usuario cuanto mayor sea el porcentaje de confianza de rechazo, mayor probabilidad tendrá de que ese sea el motivo real de rechazo.

Figura 4.4. Proceso de predicción mediante modelado.



4.6 Interfaz gráfica.

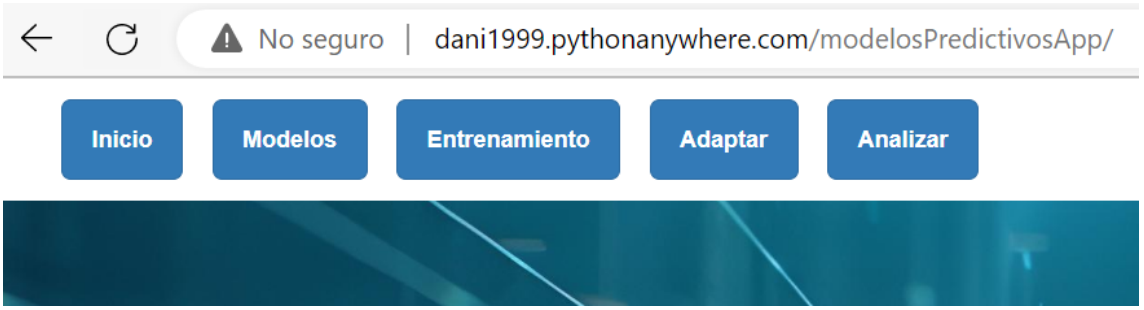
Además de lo anteriormente descrito, se ha realizado una aplicación web para el uso del modelo implementado. Se han desarrollado varias vistas con diferentes funcionalidades que se comentarán a continuación.

4.6.1 Barra de navegación

La barra de navegación es muy importante para acceder a todas las funcionalidades de la aplicación web. Se dispone de un botón “Inicio” para acceder a la página principal y de presentación de la aplicación, un botón “Modelos” para utilizar los modelos cargados en el

sistema para predecir, un botón “Entrenamiento” para entrenar futuros modelos para predecir, y finalmente un botón “Adaptar” y “Analizar” uno para tratar los datos que se quieren utilizar para entrenamiento y el otro para analizar csv y sacar información interesante de él. En la figura 4.5 se muestra como aparece en la aplicación web la barra de navegación.

Figura 4.5. Barra de navegación.



4.6.2 Pantalla de inicio

La pantalla de inicio se trata de una página de “home” en la que se presenta la aplicación web y las diferentes secciones de las que dispone. En la figura 4.6 se puede observar dicha página.

Figura 4.6. Página de inicio.

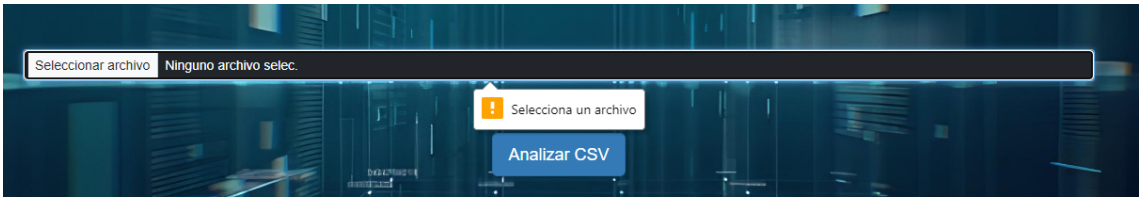
4.6.3 Pantalla de análisis de datos

La pantalla de análisis de datos se trata de una sección en la que se permite al usuario subir un fichero en formato csv para realizar un análisis de los datos que este contiene. Se observa en la figura 4.7 la vista de la misma.

Figura 4.7. Página de análisis de datos.

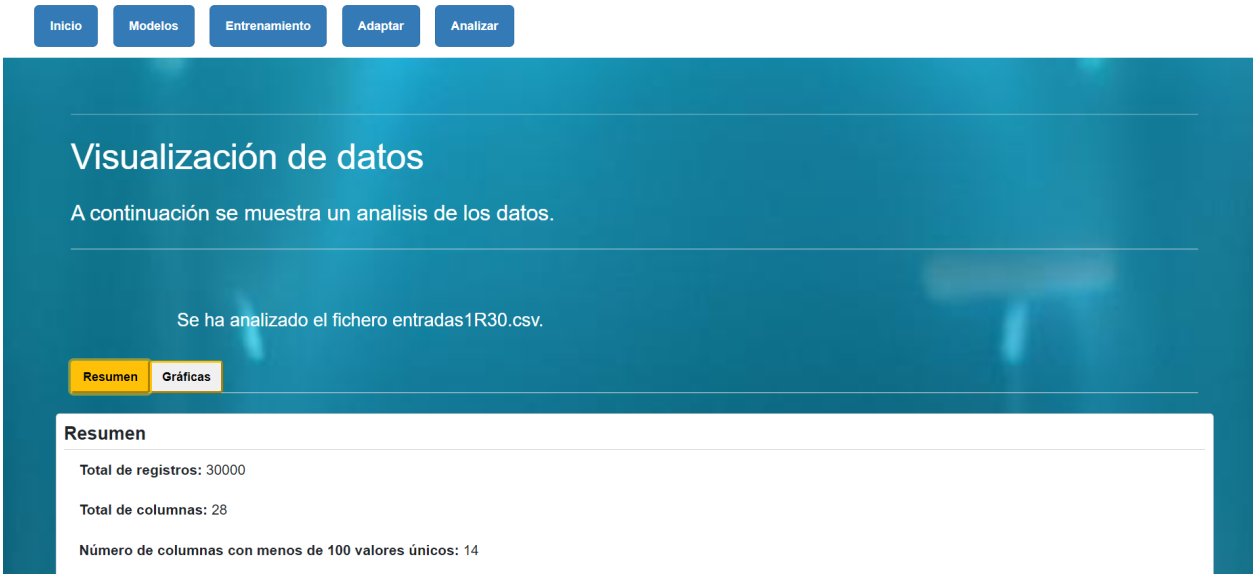
Se ha realizado una validación para asegurarnos que el usuario facilita ese fichero. En la figura 4.8 se muestra un ejemplo de como se muestra dicha validación.

Figura 4.8. Obligatoriedad de adjuntar fichero.



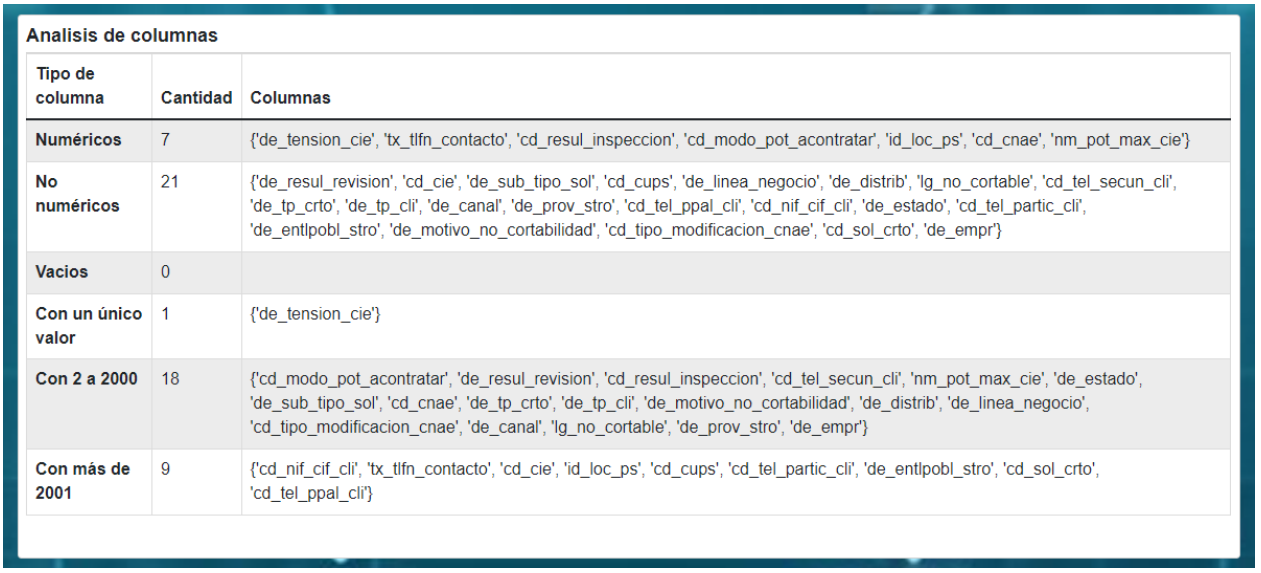
Una vez incorporado el fichero y el usuario pulse sobre el botón “Analizar CSV”, se realizará un análisis exhaustivo de esos datos. Este análisis será mostrado en una pantalla la cual tiene dos secciones, una de resumen, la cual muestra un análisis acerca del número total de registros o el número de columnas, entre otras, la otra sección es acerca de diferentes gráficas para mostrar otras informaciones referente a los datos. En la figura 4.9 se muestra dicha pantalla.

Figura 4.9.Página de visualización de datos.



En el apartado "Resumen" se muestra un análisis de columnas, en el que se indica cada columna que tipo de datos tiene y el número de columnas de ese tipo. Se observa en la figura 4.10 dicho análisis de columnas.

Figura 4.10. Análisis de columnas.

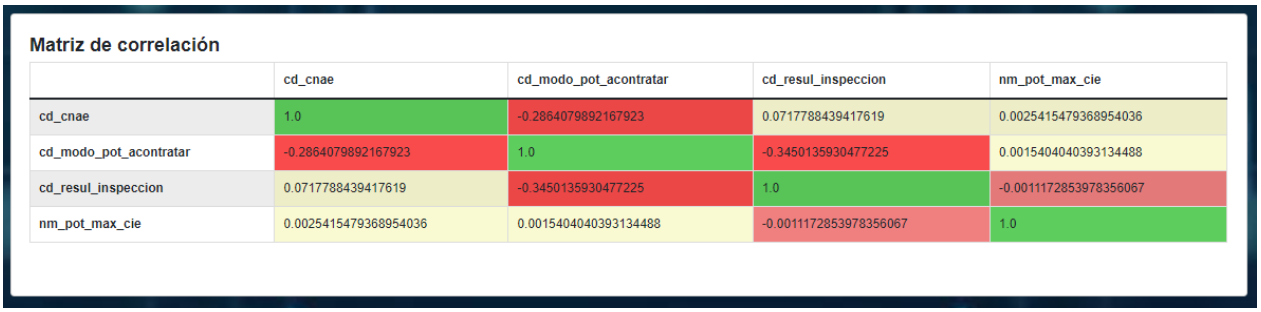


Analisis de columnas

Tipo de columna	Cantidad	Columnas
Numéricos	7	{'de_tension_cie', 'tx_tlfn_contacto', 'cd_resul_inspeccion', 'cd_modo_pot_acontratar', 'id_loc_ps', 'cd_cnae', 'nm_pot_max_cie'}
No numéricos	21	{'de_resul_revision', 'cd_cie', 'de_sub_tipo_sol', 'cd_cups', 'de_linea_negocio', 'de_distrib', 'lg_no_cortable', 'cd_tel_secun_cli', 'de_tp_crto', 'de_tp_cli', 'de_canal', 'de_prov_stro', 'cd_tel_ppal_cli', 'cd_nif_cif_cli', 'de_estado', 'cd_tel_partic_cli', 'de_entlpobl_stro', 'de_motivo_no_cortabilidad', 'cd_tipo_modificacion_cnae', 'cd_sol_crto', 'de_empr'}
Vacios	0	
Con un único valor	1	{'de_tension_cie'}
Con 2 a 2000	18	{'cd_modo_pot_acontratar', 'de_resul_revision', 'cd_resul_inspeccion', 'cd_tel_secun_cli', 'nm_pot_max_cie', 'de_estado', 'de_sub_tipo_sol', 'cd_cnae', 'de_tp_crto', 'de_tp_cli', 'de_motivo_no_cortabilidad', 'de_distrib', 'de_linea_negocio', 'cd_tipo_modificacion_cnae', 'de_canal', 'lg_no_cortable', 'de_prov_stro', 'de_empr'}
Con más de 2001	9	{'cd_nif_cif_cli', 'tx_tlfn_contacto', 'cd_cie', 'id_loc_ps', 'cd_cups', 'cd_tel_partic_cli', 'de_entlpobl_stro', 'cd_sol_crto', 'cd_tel_ppal_cli'}

Además, en el apartado de "Resumen" también se aporta una matriz de correlación entre las columnas del fichero. La matriz de correlación es útil para entender la relación entre las variables y proporciona información sobre la dependencia o independencia de las diferentes columnas. En la figura 4.11 se observa dicha matriz de correlación para un ejemplo.

Figura 4.11.Matriz de correlación.

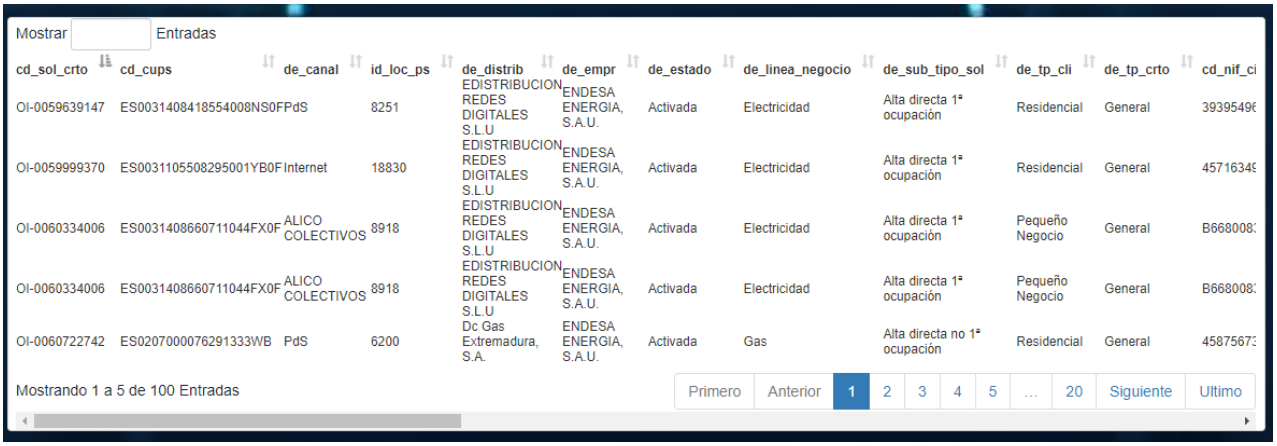


Matriz de correlación

	cd_cnae	cd_modo_pot_acontratar	cd_resul_inspeccion	nm_pot_max_cie
cd_cnae	1.0	-0.2864079892167923	0.0717788439417619	0.0025415479368954036
cd_modo_pot_acontratar	-0.2864079892167923	1.0	-0.3450135930477225	0.0015404040393134488
cd_resul_inspeccion	0.0717788439417619	-0.3450135930477225	1.0	-0.001117285397835606
nm_pot_max_cie	0.0025415479368954036	0.0015404040393134488	-0.001117285397835606	1.0

Finalmente se muestra una tabla en la que se visualizan las 100 primeras entradas del fichero para disponer de una visión general de cómo son los datos de ese fichero csv. Se puede observar un ejemplo de como se muestra dicha tabla de datos en la figura 4.12.

Figura 4.12. Tabla de datos.



Mostrar Entradas

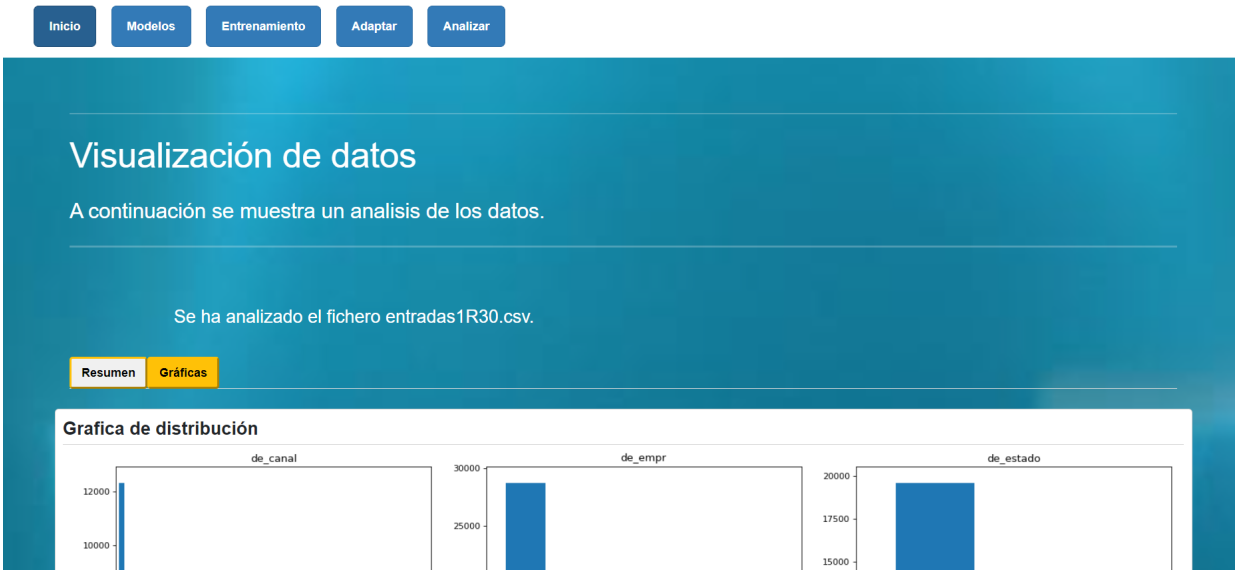
cd_sol_crto	cd_cups	de_canal	id_loc_ps	de_distrib	de_empr	de_estado	de_linea_negocio	de_sub_tipo_sol	de_tp_cli	de_tp_crto	cd_nif_ci
OI-0059639147	ES0031408418554008NS0F	PdS	8251	EDISTRIBUCION REDES DIGITALES S.L.U	ENDESA ENERGIA, S.A.U.	Activada	Electricidad	Alta directa 1ª ocupación	Residencial	General	3939549(
OI-0059999370	ES0031105508295001YB0F	Internet	18830	EDISTRIBUCION REDES DIGITALES S.L.U	ENDESA ENERGIA, S.A.U.	Activada	Electricidad	Alta directa 1ª ocupación	Residencial	General	4571634
OI-0060334006	ES0031408660711044FX0F	ALICO COLECTIVOS	8918	EDISTRIBUCION REDES DIGITALES S.L.U	ENDESA ENERGIA, S.A.U.	Activada	Electricidad	Alta directa 1ª ocupación	Pequeño Negocio	General	B668008
OI-0060334006	ES0031408660711044FX0F	ALICO COLECTIVOS	8918	EDISTRIBUCION REDES DIGITALES S.L.U	ENDESA ENERGIA, S.A.U.	Activada	Electricidad	Alta directa 1ª ocupación	Pequeño Negocio	General	B668008
OI-0060722742	ES0207000076291333WB	PdS	6200	Dc Gas Extremadura, S.A.	ENDESA ENERGIA, S.A.U.	Activada	Gas	Alta directa no 1ª ocupación	Residencial	General	4587567

Mostrando 1 a 5 de 100 Entradas

Primero Anterior 1 2 3 4 5 ... 20 Siguiente Ultimo

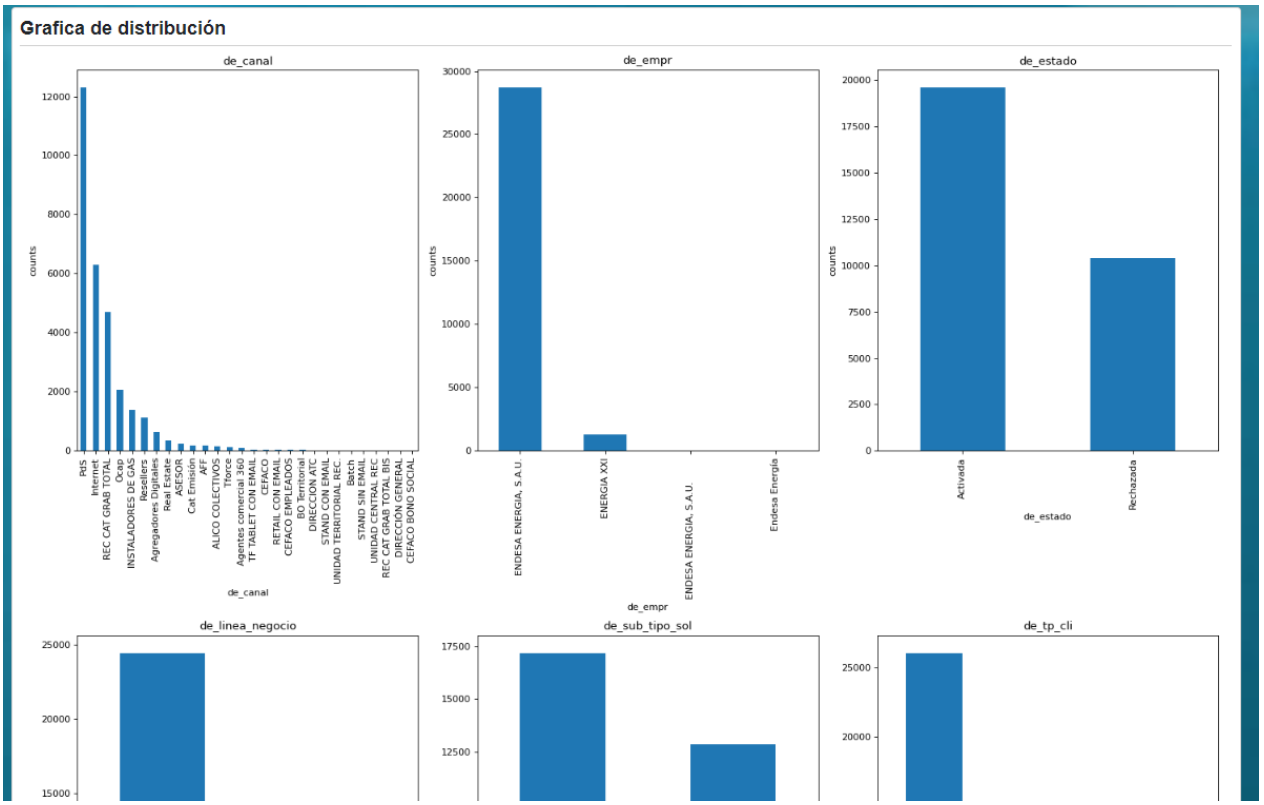
Respecto al apartado de “Gráficas”, se muestran dos tipos de gráficas de distribución y de relación de cada una de las columnas del fichero. En la figura 4.13 se observa una primera vista de la sección de gráficas.

Figura 4.13. Gráficas de visualización de datos.



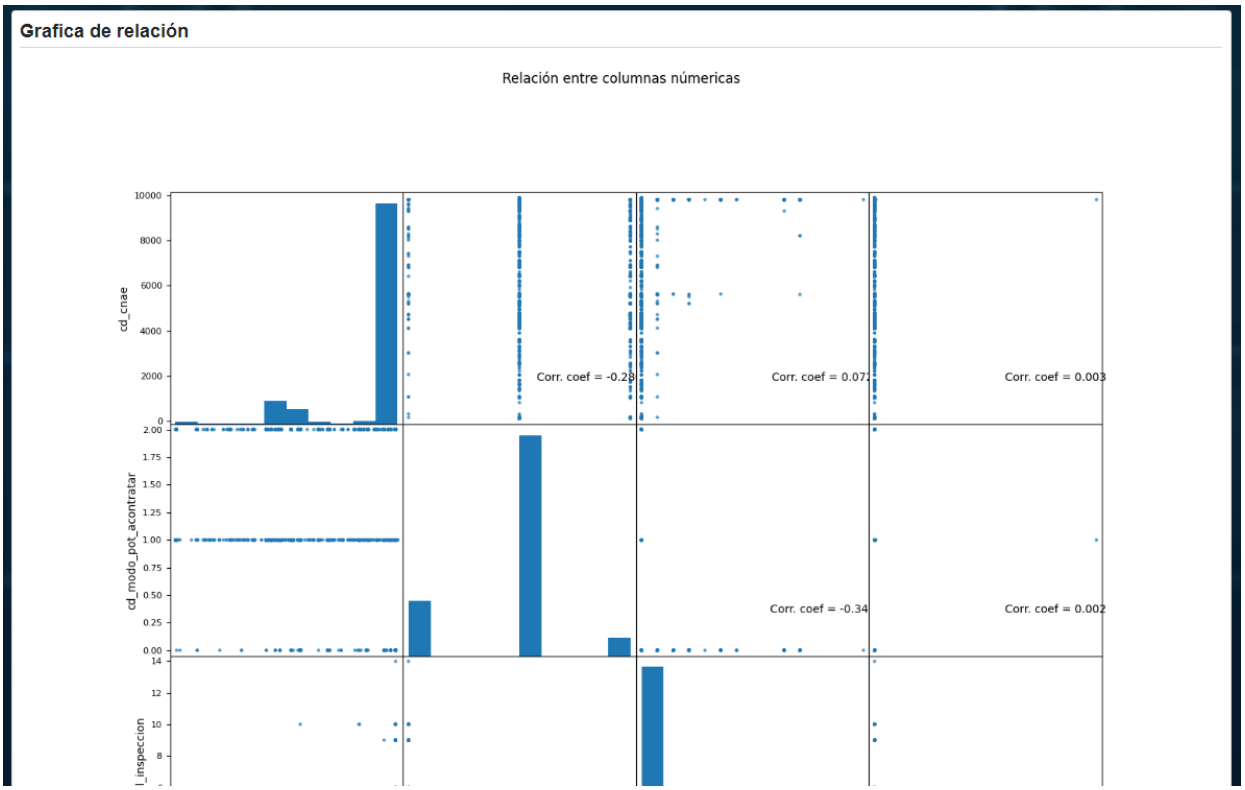
La gráfica de distribución, también conocida como histograma, es una representación visual de la distribución de los valores en una columna. Esta gráfica muestra la frecuencia o cantidad de veces que ocurre cada valor en la columna, lo que permite identificar patrones, tendencias y características de los datos. En la figura 4.14 se puede observar algunas de estas gráficas.

Figura 4.14. Gráficas de distribución de datos.



La gráfica de relación, también conocida como gráfica de dispersión o scatter plot en inglés, es una representación visual que muestra la relación o la asociación entre dos variables en un conjunto de datos. Estas son útiles para visualizar y analizar la correlación o dependencia entre dos variables, al observar la dispersión de los puntos, es posible identificar si existe una relación lineal, no lineal, positiva, negativa o ninguna relación entre las variables. Se muestran algunas de estas gráficas en la figura 4.15.

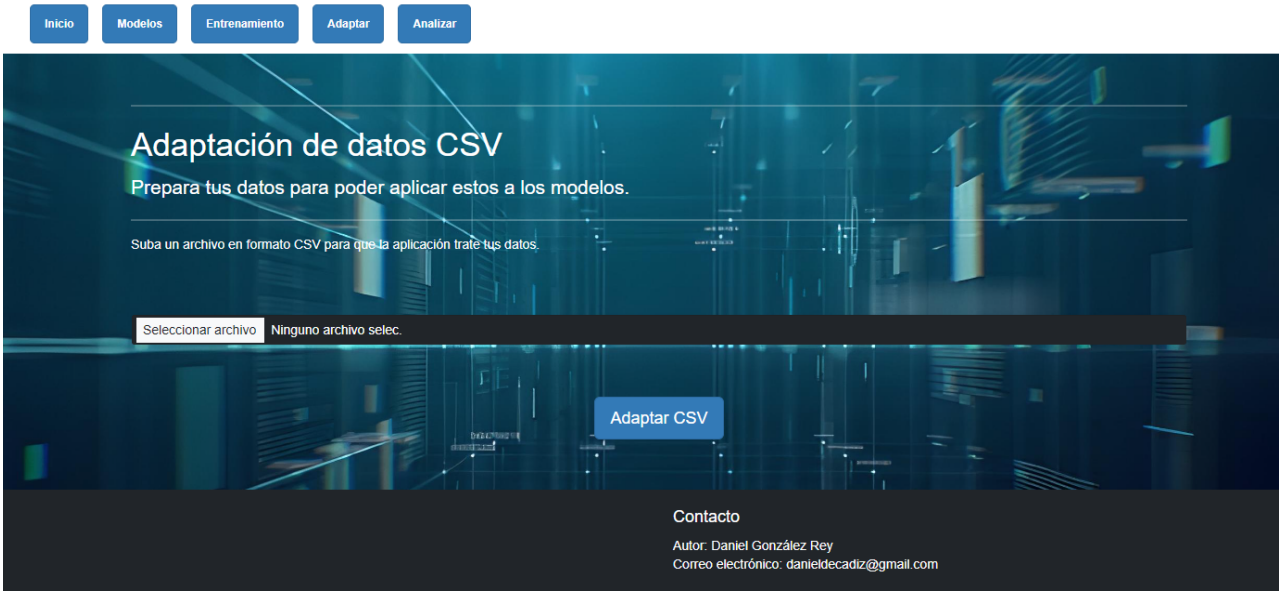
Figura 4.15. Gráficas de relación de datos.



4.6.4 Pantalla de adaptación de datos

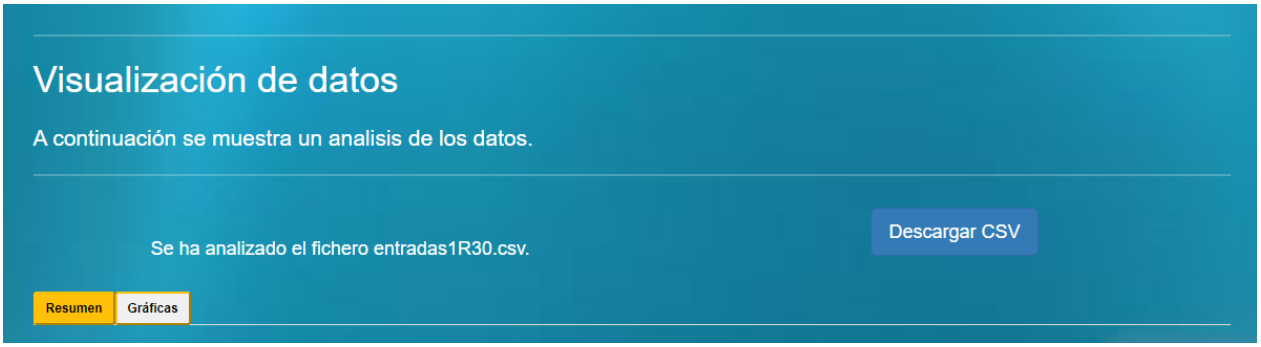
La pantalla de adaptación de datos, es una sección que es utilizada para que el usuario pueda preparar los datos para entrenar los modelos. Mediante esta vista el usuario puede facilitar el fichero en formato csv y el sistema se encargará de analizarlo y adaptarlo correctamente para su correcto entrenamiento posterior. Se observa dicha pantalla en la figura 4.16.

Figura 4.16. Página de adaptación de datos.



Una vez que el usuario suba el fichero y pulse sobre el botón ”Adaptar CSV”, el sistema tratará sus datos y facilitará un botón para descargar ese fichero csv con los datos tratado. Para descargar el fichero bastará con pulsar sobre el botón “Descargar CSV” situado bajo el encabezado en la parte izquierda. Además se le mostrará un análisis de los datos similar a la pantalla de análisis comentada anteriormente. Se puede ver el botón en cuestión en la figura 4.17.

Figura 4.17. Botón para descarga de fichero adaptado.

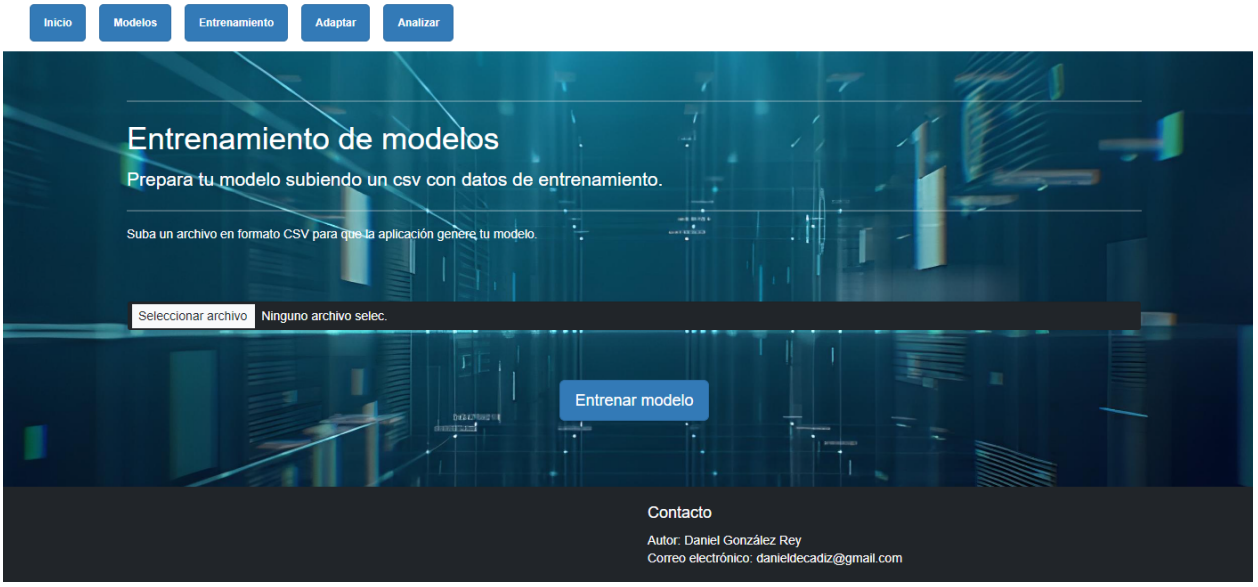


4.6.5 Pantalla de entrenamiento del modelo

La pantalla de entrenamiento es la encargada de permitirle al usuario entrenar aquellos modelos con los que posteriormente quiera predecir. El usuario únicamente debe de facilitarle los datos en un fichero csv y el sistema se encargará de crear un modelo acorde a esos datos. Se recomienda que el fichero facilitado para el entrenamiento haya pasado previamente por la sección de adaptado, para evitar que el fichero pueda no estar

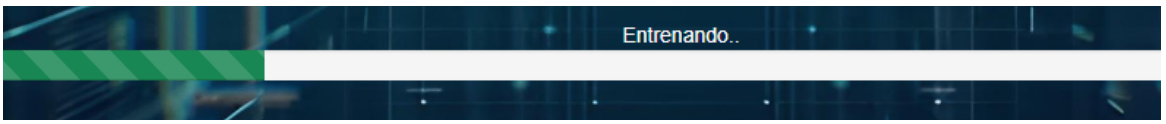
correctamente adaptado para ser entrenado. En la figura 4.18 se observa el aspecto de dicha pantalla.

Figura 4.18. Página de entrenamiento de modelos.



Debido a que el sistema puede tardar algunos minutos en entrenar el modelo en función del tamaño de este, al igual que para las secciones de análisis, adaptado y predicciones, se dispone de una barra de carga para que el usuario disponga de la información de cómo va el proceso de entrenamiento. Se muestra un ejemplo de la barra de carga en funcionamiento en la figura 4.19.

Figura 4.19. Barra de carga para entrenamiento.



Una vez que se haya finalizado el entrenamiento, ya se dispondrá del modelo cargado en el sistema, con un nombre compuesto con el nombre del fichero de entrenamiento. Además se proporciona alguna información acerca del rendimiento de este, realizado con un conjunto de test obtenido del propio conjunto de datos de entrenamiento. En la figura 4.20 se observan unos primeros resultados del modelo entrenado.

Figura 4.20. Resultados del modelo entrenado.

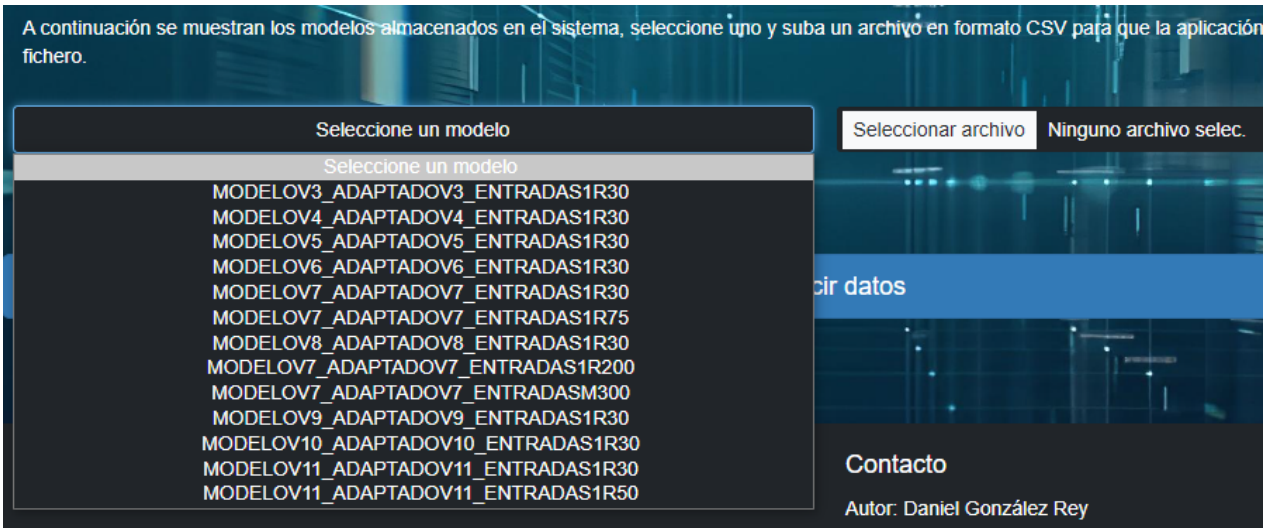
4.6.6 Pantalla de predicciones mediante modelos

Finalmente, la sección de modelos se encarga de utilizar los modelos previamente cargados para realizar las predicciones. El usuario simplemente deberá de elegir un modelo con que predecir y subir el fichero csv que se quiera utilizar para hacer las predicciones. En la figura 4.21 se observa dicha sección de modelos.

Figura 4.21. Página de predicciones .

Si se despliega el combo de modelos, se muestran todos los modelos cargados en el sistema. Se podrá elegir entre uno de ellos para realizar la predicción. Se observan unos modelos de ejemplos cargados en el sistema en la figura 4.22.

Figura 4.22. Modelos cargados en la aplicación.

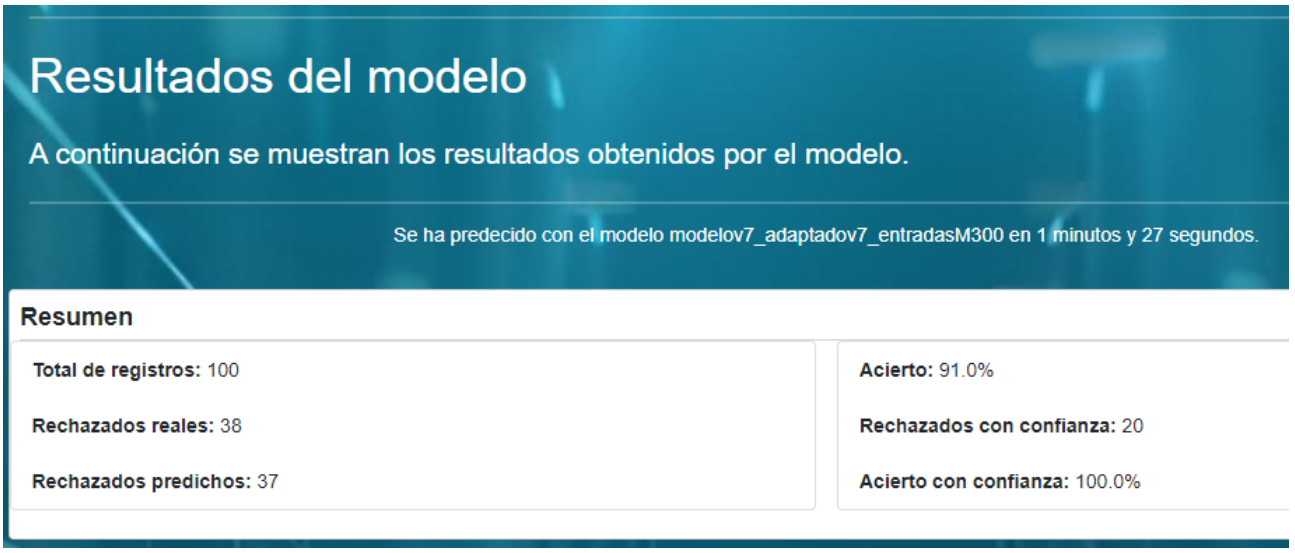


Una vez que el usuario haya elegido el modelo y subido el fichero a predecir, le aparecerá la siguiente pantalla. En ella son mostrados los resultados del modelo y finalmente un botón para descargar el fichero csv con las predicciones realizadas. En la figura 4.23 se muestra la parte superior de la pantalla de resultados.

Figura 4.23. Página de resultados de la predicción.

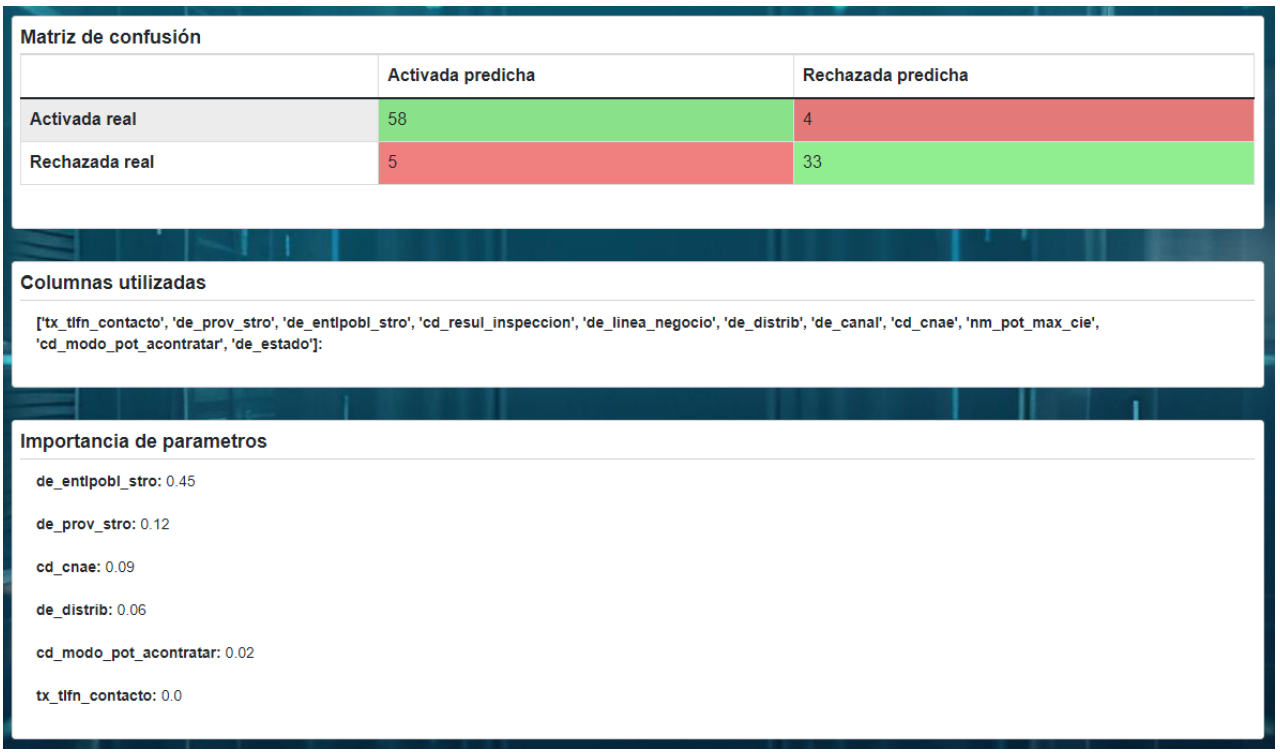
Se podrá visualizar un resumen de los resultados del modelo, mostrando el porcentaje de acierto, el número de rechazos reales frente a los predichos, o el porcentaje de aciertos con confianza mayor del 80%. Se observa en la figura 4.24 el apartado de resumen de resultados.

Figura 4.24. Resumen de resultados de la predicción.



También aparece una matriz de confusión, para mostrar un resumen de las predicciones realizadas y la efectividad del modelo. Una matriz de confusión es una herramienta utilizada en el campo del aprendizaje automático y la clasificación de datos para evaluar el rendimiento de un modelo predictivo. Además se visualizan las columnas utilizadas en el modelo y la importancia de estas para realizar las predicciones. En la figura 4.25 se puede observar como se visualiza toda esa información.

Figura 4.25. Análisis de resultados de la predicción.

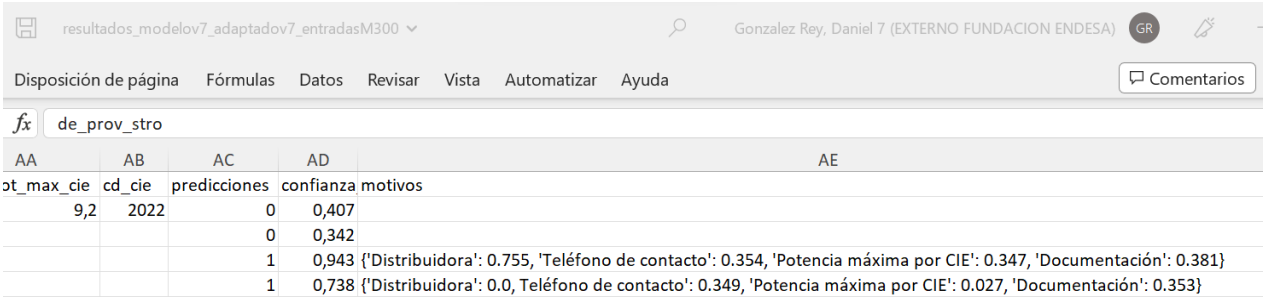


Más abajo se visualizan las 100 primeras entradas del fichero junto con sus predicciones, confianza y motivos de rechazo. Es una tabla a modo de visión general de cómo ha actuado el modelo. Finalmente se dispone del botón “Descargar CSV de resultados” para obtener el fichero csv completo con las predicciones realizadas. Se observa dicha tabla con ejemplos en la figura 4.26.

Figura 4.26. Resultados de la predicción.

A modo de ejemplo, aquí se muestra como quedaría el fichero csv resultante, queda reflejada la predicción, la confianza para cada una de ellas y sus motivos. En la figura 4.27 se puede ver como se visualiza el fichero csv resultante.

Figura 4.27. Fichero con las predicciones.



CAPÍTULO 5 - RESULTADOS Y PRUEBAS

5.1 Análisis de resultados y pruebas.

En este punto se van a analizar y comparar los resultados obtenidos con diferentes modelos y configuraciones comparados.

5.1.1 Análisis de datos importantes dentro del modelo.

El modelo predictivo desarrollado es realizado sobre el algoritmo Random Forest y sobre él se pueden realizar múltiples configuraciones. Con el modelo se realizan dos tipos de configuraciones, configuración sobre el propio algoritmo, o diferentes selección de parámetros sobre los datos.

Se comenzará viendo la diferente selección de parámetros, el número de parámetros y su importancia dependerá completamente del conjunto de datos y del objetivo que se persigue a través de ellos.

Se encuentra que el conjunto de datos principal tiene 142.302 filas y 293 columnas. Este conjunto de datos se ha obtenido accediendo a la base de datos de Endesa y obteniendo los datos desde el 1 de enero de 2022 hasta el 28 de febrero de 2023. Es importante quedarse con aquellas columnas que sean interesantes, para ello, en un primer filtrado se quedan con 83 columnas. El primer filtrado ha sido quedarse con las columnas numéricas y aquellas no numéricas pero que tengan entre 2 y 99 valores únicos diferentes. Entre las columnas que quedan se realiza una reunión con un experto en el negocio y finalmente se concreta que el conjunto de datos dispone de 18 columnas potencialmente interesantes que serían las siguientes:

- **de_canal:** Representa el canal por el que se ha presentado la solicitud.
- **de_distrib:** Indica la distribuidora a la que se le realiza la solicitud.
- **de_empr:** Indica la empresa que realiza la solicitud.
- **de_linea_negocio:** Expresa la línea de negocio a la que va referenciada esa solicitud, gas o energía.
- **de_sub_tipo_sol:** Indica el tipo de solicitud que se quiere realizar.
- **de_tp_cli:** Expresa el tipo de cliente de la que procede la solicitud.
- **de_tp_crto:** Se indica el tipo de contrato que lleva la solicitud.
- **cd_cnae:** Indica el código CNAE de la solicitud.
- **lg_no_cortable:** Se indica si la solicitud es interrumpible o no.
- **cd_modo_pot_acontratar:** Expresa el tipo de potencia a contratar en la solicitud.
- **tx_tlfn_contacto:** Se indica el número de teléfono del cliente que realiza la solicitud.
- **cd_tipo_modificacion_cnae:** Si hubiera alguna modificación de código CNAE se indicaría el tipo.

- **tiene_administrativos:** Representa si esa solicitud tiene adjunta documentación administrativa.
- **cd_resul_inspeccion:** Si ha sido necesaria realizarle a la solicitud alguna inspección se indica su resultado.
- **de_entlpobl_stro:** Indica el municipio de la que proviene esa solicitud.
- **nm_pot_max_cie:** Indica la potencia máxima a contratar debido al CIE de la solicitud.
- **tiene_tecnicos:** Representa si esa solicitud tiene adjunta documentación técnica.
- **de_prov_stro:** Indica la provincia de la que proviene esa solicitud.

Para tratar de obtener el mejor modelo posible se realizan multitud de pruebas seleccionando diferentes parámetros. En la tabla 5.1 se muestran algunos de los modelos probados junto con la importancia de los parámetros en cada uno de ellos y el porcentaje de acierto obtenido. Estos modelos están entrenados con el conjunto de datos comentado anteriormente y predicho con un fichero de 5.000 registros de los últimos 3 meses.

Tabla 5.1. Resultados de modelados.

Parámetros									**Resultados**		
de_canal	**de_distrib**	**de_empr**	**de_linea_negocio**	**de_sub_tipo_sol**	**de_tp_cli**	**de_tp_crto**	**cd_cnae**	**lg_no_cortable**	**Importancia**		**% acierto**
cd_modo_pot_acontratar	**tx_tlfn_contacto**	**cd_tipo_modificacion_cnae**	**tiene_administrativos**	**cd_resul_inspeccion**	**de_entlpobl_stro**	**nm_pot_max_cie**	**tiene_tecnicos**	**de_prov_stro**			
✓	✓	✓	✓	✓	✓	✓	✓	✓	de_entlpobl_stro	32%	**61,7%**
									de_prov_stro	15%	
✓	✓	✓	✓	✓	✓	✓	✓	✓	de_canal	12%	
✗	✓	✓	✓	✓	✓	✓	✓	✓	nm_pot_max_cie	31%	**67,5%**
									de_distrib	21%	
✓	✓	✓	✓	✓	✗	✓	✓	✗	cd_cnae	20%	

✗	✗	✓	✓	✓	✓	✓	✗	✓	tiene_tecnicos	26%	**62,2%**
									tiene_administrativos	22%	
✓	✓	✓	✓	✓	✗	✗	✓	✗	cd_modo_pot_acontratar	9%	
✗	✗	✓	✓	✓	✓	✓	✗	✓	de_linea_negocio	16%	**59,3%**
									de_empr	15%	
✓	✗	✓	✗	✓	✗	✗	✗	✗	cd_resul_inspeccion	7%	
									tx_tlfn_contacto	5%	
✗	✗	✗	✗	✓	✓	✓	✗	✓	de_sub_tipo_sol	23%	**58,6%**
									de_tp_cli	10%	
✗	✗	✓	✗	✗	✗	✗	✗	✗			
✓	✗	✗	✗	✗	✗	✗	✗	✗	de_entlpobl_stro	45%	**66,3%**
									de_canal	31%	
✗	✗	✗	✗	✗	✓	✗	✗	✓	de_prov_stro	25%	
✓	✓	✗	✗	✗	✗	✗	✓	✗	de_entlpobl_stro	34%	**64,1%**
									nm_pot_max_cie	17%	
✗	✗	✗	✗	✗	✓	✓	✗	✓	de_canal	15%	
									de_prov_stro	14%	

✗	✓	✗	✗	✗	✗	✗	✓	✗	nm_pot_max_cie	43%	**66,7%**
									de_distrib	35%	
✗	✗	✗	✗	✗	✗	✓	✗	✗	cd_cnae	22%	
✓	✓	✗	✗	✗	✗	✗	✓	✗	nm_pot_max_cie	30%	**68,0%**
									de_canal	27%	
✗	✗	✗	✗	✗	✗	✓	✗	✗	de_distrib	22%	
									cd_cnae	21%	
✓	✓	✗	✗	✗	✗	✗	✓	✗	de_entlpobl_stro	41%	**65,1%**
									nm_pot_max_cie	19%	
✗	✗	✗	✗	✗	✓	✓	✗	✗	de_canal	16%	
									de_distrib	13%	
✓	✓	✗	✗	✗	✗	✗	✓	✗	nm_pot_max_cie	29%	**67,9%**
									de_distrib	22%	
✗	✗	✗	✓	✗	✗	✓	✓	✗	de_canal	21%	
									cd_cnae	20%	
✓	✓	✗	✗	✗	✗	✗	✓	✗	nm_pot_max_cie	30%	**68,0%**
									de_distrib	22%	
✗	✓	✗	✓	✗	✗	✓	✓	✗	de_canal	21%	
									cd_cnae	19%	
✓	✓	✗	✗	✗	✗	✗	✓	✗	nm_pot_max_cie	29%	**67,4%**
									de_canal	20%	
✓	✓	✗	✓	✗	✗	✓	✓	✗	cd_cnae	20%	

									de_distrib	19%	
✓	✓	✗	✗	✗	✗	✗	✓	✗	nm_pot_max_cie	29%	**67,2%**
									de_canal	21%	
✗	✓	✓	✓	✓	✗	✓	✓	✗	de_distrib	21%	
									cd_cnae	19%	
✓	✓	✗	✗	✗	✗	✗	✓	✗	nm_pot_max_cie	30%	**68,4%**
									de_canal	26%	
✗	✓	✗	✗	✗	✗	✓	✗	✗	de_distrib	23%	
									cd_cnae	20%	
✓	✓	✗	✗	✓	✓	✗	✓	✗	nm_pot_max_cie	31%	**66,2%**
									de_canal	22%	
✗	✓	✗	✗	✗	✗	✓	✗	✗	de_distrib	21%	
									cd_cnae	20%	
✓	✓	✓	✓	✗	✗	✗	✓	✗	de_entlpobl_stro	34%	**65,0%**
									de_prov_stro	14%	
✓	✓	✗	✓	✓	✓	✓	✓	✓	de_canal	12%	
									cd_cnae	9%	
✓	✓	✗	✓	✓	✗	✗	✓	✗	nm_pot_max_cie	29%	**67,2%**
									cd_cnae	20%	
✗	✓	✗	✓	✓	✗	✓	✓	✗	de_canal	19%	
									de_distrib	19%	

Tras un exhaustivo análisis se llega a la conclusión de que los **mejores modelos encontrados son dos**. Un modelo con las columnas de 'tiene_tecnicos' y 'tiene_administrativos' y otro modelo sin ellos, el primero es llamado "mejor modelo con

documentos" y "mejor modelo sin documentos". Esta diferenciación se lleva a cabo debido a que la información acerca de si dispone de documentación o no es algo que aún se desconoce si será importante.

El mejor modelo con documentos tiene un acierto del **68%**, hay que tener en cuenta que ha sido entrenado con únicamente **142.302** por lo que previsiblemente si es entrenado con más registros mejorará su porcentaje de acierto. El mejor modelo con documentos tiene **7 columnas** que son 'de_canal', 'de_distrib', 'cd_cnae', 'tx_tlfn_contacto', 'tiene_administrativos', 'nm_pot_max_cie' y ' tiene_tecnicos'.

En cambio, el mejor modelo sin documentos tiene un acierto del **68,4%**, este modelo tiene **5 columnas** que son 'de_canal', 'de_distrib', 'cd_cnae', 'tx_tlfn_contacto' y 'nm_pot_max_cie'. A continuación se realizarán pruebas con diferentes configuraciones para tratar de mejorar resultados.

Respecto a la configuración sobre el propio algoritmo, Random Forest, al ser un algoritmo de árboles de decisión múltiples, permite configurar el número de árboles de decisión a generar. En función del número de árboles se tratará de mejorar los resultados del modelo, pero se debe de tener cuidado de no caer en un sobreajuste del mismo. En la tabla 5.2 se muestran los dos mejores modelos con las columnas utilizadas en cada uno de ellos, el número de árboles con el que han sido entrenados y el porcentaje de acierto sobre el conjunto de datos principal.

Tabla 5.2. Resultados con configuraciones .

Parámetros			**Resultados**
Modelo	**Columnas**	**Número de árboles**	**% acierto**
Mejor modelo con documentos	['de_canal', 'de_distrib', 'cd_cnae', 'tx_tlfn_contacto', tiene_administrativos, 'nm_pot_max_cie', 'tiene_tecnicos']	1.000	**68,0%**
		500	**67,9%**
		2.000	**68,1%**
		3.000	**68,1%**
		5.000	**68,1%**
		10.000	**68,1%**
		15.000	**68,0%**
Mejor modelo sin documentos	['de_canal', 'de_distrib', 'cd_cnae', 'tx_tlfn_contacto', 'nm_pot_max_cie']	1.000	**68,4%**
		500	**68,4%**

		2.000	**68,0%**
		3.000	**68,5%**
		5.000	**68,5%**
		10.000	**68,4%**

El mejor resultado para el modelo con documentos ha sido con 2.000 árboles y para el modelo sin documentos ha sido de 3.000 árboles. A partir de este momento las pruebas que se realicen serán sobre esta configuración para cada uno de ellos.

5.1.2 Tiempos de entrenamiento y predicción.

Ahora se estudiará los tiempos de procesamiento tanto para el entrenamiento del modelo como para luego las predicciones mediante el. En la tabla 5.3 se muestran los dos mejores modelos con sus columnas el número de registros para entrenamiento y predicciones y sus correspondientes tiempos y porcentajes de acierto.

Tabla 5.3. Resultados de tiempos para modelos .

Parámetros				**Resultados**		
Modelo	**Columnas**	**Nº registros a entrenar**	**Nº registros a predecir**	**Tiempos de entrenamiento (en segundos)**	**Tiempos de predicción (en segundos)**	**% acierto**
Mejor modelo con documentos	['de_canal', 'de_distrib', 'cd_cnae', 'tx_tlfn_contacto', tiene_administrativo ,'nm_pot_max_cie', 'tiene_tecnicos']	80.627	1.000	**110s**	**7s**	**68,0%**
			5.000		**13s**	**67,5%**
		140.438	1.000	**207s**	**12s**	**66,9%**
			5.000		**18s**	**67,5%**
Mejor modelo sin documentos	['de_canal', 'de_distrib', 'cd_cnae', 'tx_tlfn_contacto', 'nm_pot_max_cie']	80.627	1.000	**162s**	**9s**	**68,4%**
			5.000		**17s**	**67,3%**

		140.438	1.000	**315s**	**25s**	**69,1%**
			5.000		**32s**	**67,9%**

5.1.3 Porcentajes de acierto y confianzas

Ciertamente, el mejor resultado obtenido es muy bajo, del 68,5%. No obstante, en el modelo, realmente el porcentaje de acierto importante es aquel que se denomina aciertos con confianza. Los aciertos con confianza son aquellos que el modelo indica que la confianza de ser rechazada de esa solicitud es superior a un valor determinado por nosotros. La idea es que se logre definir **un porcentaje de confianza entre el 80% y el 99%** para ofrecer una cantidad mucho mayor de aciertos sobre los que cumplen esa condición. Por ello, sobre los mejores modelos vistos anteriormente, se observan los porcentajes de aciertos con diferentes niveles de confianza.

En la siguiente tabla 5.4 se verá, para los mejores modelos, el porcentaje de rechazos con confianza, es decir, el porcentaje de rechazos con confianza en función de los rechazos totales, el porcentaje de confianza impuesto y el porcentaje de acierto con esa confianza.

Tabla 5.4. Resultados para modelos con confianza.

Parámetros				**Resultados**
Modelo	**Columnas**	**% de rechazos con confianza**	**% de confianza**	**% acierto**
Modelo 1 (v57)	['de_canal', 'de_distrib', 'de_linea_negocio', 'de_sub_tipo_sol', 'de_tp_cli', 'cd_cnae', 'tx_tlfn_contacto', 'cd_resul_inspeccion', 'nm_pot_max_cie']	11,0%	80%	**55,4%**
		8,4%	85%	**56,7%**
		6,1%	90%	**58,3%**
		4,7%	95%	**61,9%**
		3,0%	99%	**73,6%**
Modelo 2 (v61)	['de_canal', 'de_distrib', 'de_linea_negocio', 'de_sub_tipo_sol', 'de_tp_cli', 'cd_cnae', 'tx_tlfn_contacto', 'tiene_administrativo', 'cd_resul_inspeccion', 'de_entlpobl_stro', 'nm_pot_max_cie', 'tiene_tecnico']	21,5%	80%	**62,7%**
		17,2%	85%	**65,7%**
		12,6%	90%	**69,2%**
		8,3%	95%	**69,6%**

		4,4%	99%	**69,2%**
Modelo 3 (v49)	['de_canal', 'de_distrib', 'de_linea_negocio', 'de_sub_tipo_sol', 'cd_cnae', 'tx_tlfn_contacto', 'tiene_administrativo', 'cd_resul_inspeccion', 'de_entlpobl_stro', 'nm_pot_max_cie', 'tiene_tecnico']	21,6%	80%	**66,1%**
		17,1%	85%	**69,4%**
		12.9%	90%	**68,6%**
		9,2%	95%	**75,0%**
		4,0%	99%	**76,1%**
Modelo 4 (v51)	['de_canal', 'de_distrib', 'de_linea_negocio', 'de_sub_tipo_sol', 'de_tp_cli', 'cd_cnae', 'tx_tlfn_contacto', 'tiene_administrativo', 'cd_resul_inspeccion', 'nm_pot_max_cie', 'tiene_tecnico']	12,4%	80%	**56,8%**
		9,9%	85%	**53,4%**
		7,2%	90%	**53,1%**
		5,6%	95%	**55,0%**
		3,2%	99%	**72,0%**
Modelo 5 (v40)	['de_canal', 'de_distrib', 'cd_cnae', 'tx_tlfn_contacto', 'nm_pot_max_cie']	9,5%	80%	**63,3%**
		7,7%	85%	**63,5%**
		5,8%	90%	**62,5%**
		4,6%	95%	**62,2%**
		2,6%	99%	**74,5%**
Modelo 6 (v62)	'de_canal', 'de_distrib', 'cd_cnae', 'tx_tlfn_contacto', 'tiene_administrativo', 'cd_resul_inspeccion', 'de_entlpobl_stro', 'nm_pot_max_cie', 'tiene_tecnico']	19,7%	80%	**63,1%**
		15,6%	85%	**67,1%**
		12,6%	90%	**67,9%**
		7,9%	95%	**75,2%**
		3,8%	99%	**73,1%**

Finalmente se obtiene que el mejor modelo es el modelo 3, ya que con un 99% de confianza o más, el 76,1% de los rechazos son correctos, el porcentaje más alto obtenido. Además de los rechazos totales, el 4% son rechazos con ese o más porcentaje de confianza. No obstante la elección del porcentaje de confianza será algo que seleccione el departamento a necesidad del negocio, en vistas de obtener el mayor porcentaje de ahorro de trabajo. Lo que se define en este apartado es que el modelo 3 es el que mejores resultados alcanza, ya que

obtiene el mejor porcentaje de acierto y respecto al promedio entre una confianza del 80% al 99% también obtiene el mejor porcentaje de acierto, un 71,04%.

5.2 Resultados del modelo predictivo.

Una vez obtenido el modelo óptimo se implementará en la aplicación web para utilizarlo. Además de los resultados anteriormente destacados, hay que tener en cuenta que el modelo también ofrece los posibles motivos de rechazo para aquellas que tengan el 95% de confianza o más de serlo, a fin de ayudar a la corrección de las mismas.

En cada uno de los motivos definidos en el modelo, se le muestra al usuario el porcentaje de confianza de rechazo para cada uno de los modelos, de esta forma aquellos que tengan un porcentaje mayor querrá decir que tienen más probabilidad de ser el motivo de rechazo.

En la tabla 5.5, se muestran algunas predicciones realizadas por el modelo a modo de ejemplo de cómo funciona el sistema. En ella aparecen algunos datos propios de la solicitud como 'nm_pedido', 'de_canal', 'cd_cnae' o 'nm_por_max_cie', además de la predicción realizada, la confianza de ser rechaza y si corresponde los motivos de rechazo.

Tabla 5.5. Ejemplos de resultados con mejor modelo.

nm_pedido	de_canal	cd_cnae	nm_pot_max_cie	Predicción	Confianza	Motivos
199616055	PdS	6910	27.71	Activada	0,153	
199623538	PdS	9820	5.75	Rechazada	0.986	{'Distribuidora': 0.876, 'Teléfono de contacto': 0.354, 'Documentación': 0.352, Potencia máxima CIE': 0.347}
191890425	Internet	9820		Rechazada	0,866	
198436436	Agregadores Digitales	5210	5.75	Rechazada	0,642	
199258995	Internet	9820	57.2	Activada	0,271	
198459375	Ocap	4729	9.2	Rechazada	0,978	{'Distribuidora': 0.337, 'Teléfono de contacto': 0.775, 'Documentación': 0.428, Potencia máxima CIE': 0.347}

5.3 Presupuesto y análisis financiero.

Para calcular el presupuesto necesario para el proyecto se tendrá en cuenta tanto los recursos humanos como los materiales. Respecto a los recursos humanos se necesitan diferentes roles para los diferentes desarrollos necesarios y deberá dedicar un número diferente de horas. Respecto a los recursos materiales se ha estimado la necesidad de un ordenador de alta capacidad para poder probar los modelos antes de ser enviados a producción, la licencia de Office 365 para poder realizar análisis de datos y documentar el proyecto, y contratar el servicio de hosting donde alojar la aplicación. Respecto al hosting hay multitud de empresas en el mercado pero se ha elegido PythonAnywhere debido a que permite un plan inicial para realizar las pruebas necesarias sin costo. Se ha detallado toda la información referente al presupuesto en la tabla 5.6 dando como resultado un coste total de 6.869€.

Tabla 5.6. Costes del proyecto.

Recursos humanos			
Rol	**Precio por hora**	**Horas totales**	**Coste**
Ingeniero de datos	18€	175	3.150€
Desarrollador web	12€	75	900€
Arquitecto de sistemas	18€	50	900€
Tester	12€	50	600€
			5.550€
Materiales			
Material			**Coste**
Ordenador con altas capacidades de cómputo			1.200€
Hosting en PythonAnywhere			Gratuito
Microsoft Office 365			69€
			1.269€
Costes totales			
			6.819€

5.4 Tabla de ahorros.

El sistema trata de ahorrar gestiones al departamento de rechazos de la empresa, lo que produce un ahorro en los costes. El departamento de rechazos envía una solicitud a la distribuidora, si es rechazada dicha solicitud, entonces el departamento de rechazos deberá dedicar un tiempo en corregir esa solicitud y volver a enviarla a la distribuidora.

El ahorro se produce al evitar el rechazo de una solicitud. Si una solicitud es rechazada, se generará un costo adicional al tener que reenviar la solicitud a la distribuidora y alargar el tiempo de gestión de esa solicitud. Además, revisar el rechazo a posteriori tiene un gasto extra ya que duplica el trabajo de otros departamentos internos de la empresa que están entre el departamento de rechazos y la distribuidora.

Este ahorro dependerá de la confianza de rechazo a partir de la cual se revisarán las solicitudes para evitar los rechazos. Cuanto mayor sea la confianza de rechazo menos solicitudes serán revisadas y por tanto menor ahorro, aunque hay que tener en cuenta que también mayor será la probabilidad de error por parte del modelo.

En la tabla 5.7 se indica un ejemplo de las solicitudes que se ahorra de gestionar el equipo en un caso concreto. En el caso de una confianza de rechazo del 85% hay 34 solicitudes que con esa o más confianza el modelo detecta que van a venir rechazadas por la distribuidora. El porcentaje de acierto para esa confianza es del 69,4% por lo que 24 solicitudes serán rechazos reales que el equipo podrá evitar, y 10 solicitudes el equipo las revisará y verá que realmente son correctas y serán activadas, es decir, enviadas sin modificación. Por tanto el ahorro real en este caso, sería el coste que supondría el rechazo de esas 24, aunque habría que añadirle el coste de revisar 10 extras. El coste de un rechazo es de aproximadamente el doble que el de revisión por lo que el ahorro sería de 14 solicitudes. Se expresa el ahorro en la siguiente fórmula: $34x - 24(2x) = 14x$, donde x hace referencia a el costo de revisión, por lo que sí nos gastamos 34x pero el costo sin ello hubiera sido 24 por el doble de x entonces el ahorro es de 14x. En la tabla se muestra un ejemplo para diferentes porcentajes de confianza de rechazo.

Tabla 5.7. Ejemplo de ahorro en función de la confianza de rechazo.

Confianza de rechazo	Rechazos detectados	Porcentaje de acierto	Solicitudes rechazadas evitadas	Solicitudes revisadas añadidas	Ahorro
80%	42	66,1%	28	14	14
85%	34	69,4%	24	10	14
90%	26	68,6%	18	8	10
95%	18	75,0%	14	4	10
99%	8	76,1%	6	2	4

Una vez visto el funcionamiento para casos concretos, en la tabla 5.8 se muestra la misma tabla de ahorro pero con porcentajes para tratar de obtener una estimación de ahorro. Se observa el porcentaje de solicitudes que se ahorran de gestionar, así como aquellas que se revisan sin ser necesario debido al porcentaje de acierto del modelo.

Tabla 5.8. Ahorro en función de la confianza de rechazo.

Confianza de rechazo	Porcentaje de rechazos detectados	Porcentaje de acierto	Porcentaje de solicitudes rechazadas evitadas	Porcentaje de solicitudes revisadas añadidas	Porcentaje de ahorro
80%	21,6%	66,1%	14,3%	7,3%	7,0%
85%	17,1%	69,4%	11,9%	5,2%	6,7%
90%	12,9%	68,6%	8,8%	4,1%	4,7%
95%	9,2%	75,0%	6,9%	2,3%	4,6%
99%	4,0%	76,1%	3,1%	0,9%	2,0%

CAPÍTULO 6 - CONCLUSIONES Y TRABAJOS FUTUROS

6.1 Conclusiones.

Tras el desarrollo de este Trabajo Fin de Máster se ha obtenido una herramienta software capaz de ahorrar tiempo y dinero a la empresa, haciendo uso de la inteligencia artificial y el Machine Learning.

El contexto a predecir por el modelo es uno muy complejo, con muchas casuísticas diferentes y muchas de ellas no expresadas en los datos. Esta herramienta es capaz de predecir con un acierto del 76,1% en un contexto concreto, aunque también hay que tener en cuenta que en muchos otros contextos el acierto es más bajo. No obstante, más allá del porcentaje de acierto hay que tener en cuenta la utilidad de este software para mejorar los procesos de la empresa. Estos modelos siempre podrán ir siendo mejorados para obtener mayores ahorros y ventajas.

Es importante destacar que este proyecto ha sido un trabajo conjunto, entre el equipo de Endesa y el equipo de la universidad, lo cual demuestra que la cooperación entre empresas y universidad es vital para la mejora de procesos dentro de la sociedad y el avance de esta.

6.2 Trabajos futuros.

Respecto a los trabajos futuros hay multitud de ellos, desde la mejora del propio modelo o de la integración de este, hasta llevar este tipo de tecnología a otros ámbitos.

El modelo desarrollado durante este Trabajo de Fin de Máster tiene bastante margen de mejora. Por una parte se podría tratar de mejorar la ingeniería de características realizada sobre las columnas, tratando de obtener mayor información de las columnas ya utilizadas. También se podría explorar nuevamente la base de datos de la compañía en busca de nuevas columnas que aporten más información al modelo. Otra mejora posible es la integración del modelo en servidores especializados, de manera que pueda entrenar modelos más grandes y potentes.

Por otro lado, esta tecnología se puede llevar a multitud de ámbitos más, desde otros departamentos de Endesa, como facturación o calidad, los cuales verían una mejora y eficiencia en sus procesos, hasta en otras empresas de diferentes sectores. La inteligencia artificial puede ser aplicada por ejemplo, al sector de la generación, permitiendo a los generadores predecir cuánta energía será necesaria producir. También a otros sectores fuera del sector energético como puede ser el médico, para análisis médicos, o el sector servicios, para predecir tendencias de consumo.

CAPÍTULO 7 - BIBLIOGRAFÍA

1. Endesa. (2022). Quienes somos.
2. Adrian, H., & MJ, K. (2007). El libro de Django 1.0.
3. Universidad de Alcalá. (s.f). Scikit-learn, herramienta básica para el data science en Python.
4. Centeno Martín-Romero, A. (2020). Big Data. Técnicas de machine learning para la creación de modelos predictivos para empresas.
5. Microsoft Azure. (s.f.). ¿Qué es la inteligencia artificial?.
6. Microsoft Azure. (s.f.). ¿Qué es el aprendizaje automático?.
7. Holovaty, A., & Kaplan-Moss, J. (2009). The definitive guide to Django: Web development done right.
8. Upbe. (s.f.). Aplicación de la IA en los negocios: casos prácticos por sectores.
9. Deloitte. (s.f.). Impacto de la IA en las empresas.
10. Jobted. (s. f.). ¿Cuánto Cobra un Ingeniero de Software? (Sueldo 2023).

Resumen

El presente documento detalla el desarrollo y la implementación de un modelo predictivo aplicando técnicas de Machine Learning y ciencia de datos, así como una aplicación web para su uso. Este modelo está destinado a predecir el resultado de solicitudes dentro del departamento de gestión de rechazos de Endesa, empresa encuadrada en el sector energético.
Durante el desarrollo de este trabajo se han definido una serie de bases teóricas y técnicas, necesarias para dicho trabajo. Además se ha detallado exhaustivamente la manera de implementar dicho modelo, así como la integración de la aplicación web. Finalmente se ha compartido un análisis de los resultados obtenidos y una conclusión acerca de los mismos y del trabajo en general.
En conclusión, se han conseguido alcanzar los objetivos propuestos, se ha logrado obtener un modelo predictivo de calidad integrado en una aplicación web útil y cómoda para el uso de los empleados. De esta manera se ha aportado un salto en la transformación digital de la empresa y en la automatización de procesos, que mejoran la calidad y eficiencia de los servicios.

Palabras clave: Modelo predictivo, machine learning, ciencia de datos, aplicación web, automatización, transformación digital, sector energético.

Abstract

This document provides a detailed account of the development and implementation of a predictive model using Machine Learning and data science techniques, along with a web application for its utilization. This model is designed to predict the outcome of requests within the rejection management department of Endesa, a company operating in the energy sector.
Throughout the course of this work, a set of theoretical and technical foundations necessary for the project have been defined. Additionally, the implementation of the model and the integration of the web application have been thoroughly described. Finally, an analysis of the obtained results has been presented, along with a conclusion regarding these outcomes and the overall work.
In conclusion, the proposed objectives have been successfully achieved. A high-quality predictive model integrated into an easy-to-use web application for employee convenience has been developed. By doing so, a significant contribution to the company's digital transformation and process automation has been achieved, improving the quality and efficiency of the services provided.

Key words: Predictive model, machine learning, data science, web application, automation, digital transformation, energy sector.