Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Ingeniería Informática

Trabajo Fin de Grado

Cálculo de Primas de Seguros Generales mediante el Uso de Algoritmos de Regresión Lineal y Series Temporales

Autor: Cristina Fernández-Simal Bernard

Tutor: Vicente Martínez Orga

Madrid, junio 2023

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Ingeniería Informática
Título: Cálculo de Primas de Seguros Generales mediante el Uso de Algoritmos de Regresión Lineal y Series Temporales
Junio 2023

Autor: Cristina Fernández-Simal Bernard

Tutor:
Vicente Martínez Orga
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Agradecimientos

Para empezar, quiero dar las gracias a mi tutor por su gran ayuda durante estos meses.

También quiero agradecer a mi familia por su apoyo incondicional durante estos últimos cinco años y por haber creído siempre en mí.

Gracias especialmente a mi hermana, empezamos este camino juntas y no podría estar más orgullosa de lo que hemos conseguido.

Gracias a Fati, la carrera no habría sido igual sin ti.

Por último, gracias a todos mis compañeros de clase por haber hecho de esta experiencia algo inolvidable.

Resumen

En este Trabajo de Fin de Grado (TFG) se pretende calcular primas de seguros generales aplicando técnicas de Machine Learning. El sector asegurador es un sector clave de la economía en cualquier país. Los seguros juegan un papel fundamental en la sociedad, al proteger a los asegurados, que bien pueden ser particulares o empresas, contra posibles riesgos. Y es que la contratación de una póliza de seguros aporta al asegurado beneficios tales como una mayor seguridad y tranquilidad. Además, también permiten reducir las pérdidas y riesgos asociados a diferentes factores, según lo que se desee cubrir con dicho contrato (salud, automóvil, hogar, etc.).

Cuando se contrata una póliza de seguro, el asegurado se compromete a pagar, a cambio de la cobertura, un precio que se denomina la prima del seguro. El pago de esta prima debe de ser total antes de que el contratante pueda beneficiarse de las coberturas que ofrece la póliza. Además, se puede fraccionar en pagos anuales, semestrales, trimestrales, etc., según acuerden las partes firmantes. Calcular esta prima de seguro es una operación compleja, ya que implica el cálculo de probabilidades. Efectivamente, el importe de esta viene determinado por el riesgo asociado al siniestro que se pretende cubrir. De esta forma, la precisión de esta probabilidad determinará un mayor o menor ajuste en el precio de la prima.

Varios estudios han demostrado que algunos algoritmos de predicción basados en Machine Learning pueden ser muy útiles a la hora de predecir con una mayor precisión la ocurrencia de un acontecimiento dado, los valores que puede tomar cierta variable, etc. Este trabajo se centra en la aplicación de algoritmos de Regresión Lineal, basados en series temporales, con el objetivo de conseguir la mejor predicción posible acerca de la prima de seguro que deberá de pagar un asegurado, es decir, aquella con el menor margen de error. Dicho importe se comparará con importes reales (que proveen distintos calculadores de primas de seguros) para poder realizar la comparativa entre ambos y sacar las conclusiones pertinentes. De esta forma, tras el desarrollo del proyecto se ha podido observar que el modelo de Regresión Lineal presenta una precisión bastante alta, no obstante, existe una mayor diferencia entre los precios predichos y aquellos de primas reales a partir de altos importes.

Palabras clave: Machine Learning, prima de seguro, Regresión Lineal, seguros de hogar, riesgo, siniestro.

Abstract

The following Final Degree Project (FDP) consists of calculating insurance premiums using Machine Learning algorithms. The insurance sector is key to every country's economy. Insurance policies play a fundamental role in society, allowing protection for the policyholder, who may be a person or a business, against uncertainties. Thus, the insured can benefit from more safety and tranquillity, as well as a reduction in the possible loss and risks associated with a certain event or activity (health, automobile, home, etc.).

When taking out insurance, the policyholder guarantees a periodical payment (that can be yearly, semestral, trimestral, monthly, etc.) in exchange for a coverage. This payment is known as the insurance premium. Calculating this price is complex, as it involves statistics and probability. As the nature of the product is giving protection against uncertain events, the price of the premium is based on the risk associated with those events (i.e., the probability of them happening). Thus, the precision of this probability will affect the price of the premium.

Some studies have proven that prediction algorithms based on Machine Learning techniques can be a powerful tool for more accurate predictions. Therefore, this project will focus on the use of Linear Regression based on Time Series, to obtain the more accurate premium associated with a policy. Hence, the different outcomes obtained will be compared with one another, as well as with the results from an online insurance premium calculator. After developing the model, it has been concluded that the precision offered is appropriate. Nonetheless, the accuracy of the algorithm compared to real insurance premium values decreases with rising prices.

Key words: Machine Learning, insurance premium, Linear Regression, home insurance, risk, casualty.

Tabla de contenidos

Índice de Figuras

Índice de Ecuaciones

Índice de Gráficos

Índice de Tablas

Índice de Ilustraciones

1 Introducción

El sector asegurador supone alrededor del **5% del PIB español** [1]. Se trata de un sector clave para la economía, aportando protección y seguridad ante imprevistos. Los seguros abarcan un gran abanico de actividades, por lo que se dividen en una gran diversidad de tipos. No obstante, se suelen distinguir dos grandes categorías de seguros: los seguros de vida y los seguros generales o de no vida. En este trabajo se pone el foco en estos últimos.

Según la definición de la Fundación Mapfre, el seguro de vida "es uno de los tipos del seguro de personas en el que el pago por el asegurador de la cantidad estipulada en el contrato se hace depender del fallecimiento o supervivencia del asegurado en una época determinada." [2]. En contraposición se encuentran los **seguros generales**. Allianz los define como "el seguro directo distinto del seguro de vida, a la denominación de la autorización que permite operar a las compañías de seguros en todos los ramos no vida." [3]. Este tipo de seguros pretende cubrir la pérdida en el patrimonio del asegurado. Dentro de los seguros generales, encontramos varios tipos de seguros, divididos por ramos. Un ramo se podría definir como un "conjunto de modalidades de seguro relativas a riesgos de características o naturaleza semejantes" [4]. De esta forma los ramos son los siguientes [3]:

- Automóvil.
- Salud (Enfermedad).
- Hogar.
- Vehículos terrestres (no ferroviarios).
- Vehículos ferroviarios.
- Vehículos aéreos.
- Vehículos marítimos, lacustres[1] y fluviales.
- Mercancías transportadas (comprendidos los equipajes y demás bienes transportados).
- Incendio y elementos naturales.
- Otros daños a los bienes.
- Responsabilidad civil[2] en vehículos terrestres automóviles.
- Responsabilidad civil en vehículos aéreos.
- Responsabilidad civil en vehículos marítimos, lacustres y fluviales.
- Responsabilidad civil en general.
- Crédito.
- Caución[3] (directa o indirecta).

---

[1] El transporte lacustre consiste en la navegación que realizan embarcaciones a través de los lagos o canales, movilizando carga y/o pasajeros entre dos o más puertos ubicados en las riberas de estos lagos, uniendo puntos geográficos diferentes en el ámbito nacional e internacional.

[2] La Responsabilidad Civil es la obligación que tiene una persona física o jurídica (sociedad o administración pública) de reparar o compensar por los daños y perjuicios que ocasione sobre otra persona, su patrimonio o sus bienes, generalmente mediante una indemnización.

[3] La caución es la garantía que se entrega con el fin de asegurar que se cumplirá con lo pactado o prometido.

- Pérdidas pecuniarias diversas[4].
- Defensa jurídica.
- Asistencia.
- Decesos.

Por tanto, al existir tantos tipos diferentes de seguros dentro del mismo ramo, el cálculo de la prima se va a centrar en uno de ellos: los **seguros de hogar**. De esta forma, el objetivo es facilitar la obtención de datos, así como homogeneizar el posterior desarrollo del modelo.

Así, se pueden definir los siguientes objetivos:

- Entender cómo se calculan las primas de seguros de hogar.
- Entender cómo funcionan los algoritmos de Regresión Lineal y cómo usarlos para el cálculo de primas de seguros de hogar.
- Buscar y tratar los datos necesarios para realizar el cálculo y alimentar el modelo.
- Analizar los resultados obtenidos y contrastarlos con resultados reales.

[4] Son aquellos seguros en los que la compañía de seguros indemniza al asegurado por la pérdida del rendimiento económico que hubiera podido alcanzar en un acto o actividad, de no haberse producido el siniestro descrito en el contrato.

2 Estado del arte

Para llevar a cabo este proyecto, se van a emplear una serie de herramientas que se van a detallar a continuación. Además, tras explicarlas, así como la razón por la cual se ha decidido emplear cada una de ellas, se explicará la motivación detrás de la elección de este trabajo.

2.1 Herramientas empleadas

2.1.1 Algoritmo de Regresión Lineal

Este trabajo se basa en el cálculo de primas de seguros generales usando un algoritmo de Regresión Lineal. Es un tipo de algoritmo de aprendizaje supervisado que se emplea para Machine Learning y Estadística. Según IBM, el **aprendizaje supervisado** "es una subcategoría del Machine Learning y la inteligencia artificial. Se define por el uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifiquen datos o predigan resultados de forma precisa. A medida que los datos se introducen en el modelo, este ajusta sus ponderaciones hasta que dicho modelo se haya ajustado adecuadamente, lo que ocurre como parte del proceso de validación cruzada." [5]

La regresión lineal se emplea para **analizar** una serie de datos y **predecir** valores desconocidos, normalmente valores futuros, que puede tener. Para ello, hace uso de otro valor de datos que tiene relación con los anteriores y es conocido [6]. De esta forma, tiene su base en un modelo matemático donde la variable que queremos predecir (variable dependiente) y aquella conocida (variable independiente) se relacionan mediante una ecuación lineal. La variable dependiente es única, mientras que se pueden tener varias variables independientes.

Se dice que este algoritmo es de **naturaleza paramétrica** [7] ya que realiza predicciones o suposiciones acerca de un conjunto de datos. De esta forma, si este se comporta como indican las suposiciones, la regresión aporta resultados fiables. En caso contrario, la precisión del modelo resulta más dudosa.

El algoritmo se basa en dos tipos de variables. Por una parte, la **variable dependiente** (variable Y), también llamada variable objetivo, es aquella que, como su nombre indica, depende de las variables independientes. Se trata de aquello que se intenta predecir. Por otra parte, las **variables independientes** o explicativas o predictivas (variable X) son aquellas que sufren modificaciones con el objetivo de averiguar los posibles valores que puede tomar una variable objetivo [7].

La regresión lineal en el Machine Learning debe de cumplir los siguientes cuatro requisitos [6]:

- **Relación Lineal**: tiene que existir una relación lineal entre las variables dependientes e independientes. Una forma de comprobarlo es trazar una gráfica de dispersión y comprobar si los datos se sitúan a lo largo de una línea recta.
- **Independencia residual**: para evaluar la precisión de una predicción se usan los residuos. Se trata de la diferencia que existe entre el valor

de los datos conocidos y aquel predicho. En este caso, se busca que no exista un patrón entre los residuos.

- **Normalidad**: los residuos, además de ser independientes entre sí también deben de seguir una distribución normal. Se puede analizar su normalidad representándolos en una gráfica Q-Q, por ejemplo. En el caso en el que estos no estén normalizados, se pueden transformar, por ejemplo, eliminando los valores atípicos.
- **Homocedasticidad**: se presupone que los residuos varían de forma constante. Si este no es el caso, puede provocar una falta de fiabilidad en los resultados obtenidos.

La ecuación de regresión lineal varía según se trate de regresión lineal simple o múltiple. La regresión lineal simple se usa cuando únicamente se tiene una variable independiente, mientras que la múltiple es útil cuando se dispone de varias variables independientes. Ambas se definen mediante una función lineal.

La función lineal que sigue una **Regresión Lineal Simple** es la siguiente:

$$y = ax + b$$

Ecuación 1

Donde y representa la variable dependiente y x la independiente. Además, a representa la pendiente. Se trata de una magnitud de cambio que pasa por y cuando x varía. La variable b es una constante, conocida como intercepto ya que $y = b$ cuando x es igual a 0 [7].

Otra forma de representar la función de Regresión Lineal Simple es la siguiente:

$$Y = \beta 0 * X + \beta 1 + \varepsilon$$

Ecuación 2

En este caso β0 y β1 son dos constantes desconocidas que representan la pendiente de regresión. Por su parte, ε es el término de error [6].

De esta forma, en ambas representaciones observamos una sola variable X independiente.

La función lineal que sigue una **Regresión Lineal Múltiple** es la siguiente [7]:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b$$

Ecuación 3

En este caso, tenemos n variables independientes. Esta función también se puede representar de la siguiente manera:

$$Y = \beta 0 x0 + \beta 1 x1 + \beta 2 x2 + \cdots + \beta N x N + \varepsilon$$

Ecuación 4

Se puede observar que, a medida que aumentan las variables independientes, aumentan las constantes β que representan los pesos de las

variables independientes. Estas constantes miden, en realidad, el efecto que tienen cada una de las variables x sobre la variable dependiente y [6].

2.1.2 Cálculo de primas de seguros generales

La **prima** de un seguro representa el importe que el tomador de cierta póliza de seguro debe de pagar a la aseguradora en compensación por los servicios prestados. Y es que, al firmar un contrato de seguro, la compañía aseguradora está asumiendo un **riesgo** basado en la protección económica del patrimonio de una persona física o jurídica (en este caso de los seguros generales, en el caso de los seguros de vida se habla del propio asegurado) ante **acontecimientos futuros imprevistos y desconocidos**. De esta forma, la asunción del riesgo reside en la incertidumbre asociada a los productos y servicios que ofrece la sociedad.

La determinación del precio que deberá de pagar el asegurado para compensar este riesgo es, por tanto, de suma importancia de cara a asegurar la rentabilidad y durabilidad del negocio. Este cálculo se estudia en el Trabajo de Fin de Grado de la rama de Administración y Dirección de Empresas (ADE) homólogo titulado **Análisis sobre el Cálculo de Primas de Seguros Generales** [8]. Efectivamente se trata de un trabajo que, en el marco de este proyecto, sirve de prefacio para entender el objeto principal de estudio.

En este apartado se van a explicar, de forma resumida, los puntos principales expuestos. Los libros acerca de matemática actuarial, ciencia en la que se basa la tarificación en el sector asegurador, determinan que la prima de un seguro está formada por la suma de los siguientes elementos:

- **Siniestralidad**: se trata del coste derivado de los siniestros producidos y para los que se ofrece cobertura en la póliza[5].
- **Gastos**: se deben tener en cuenta tanto los gastos fijos[6] como los gastos variables[7] a los que tiene que hacer frente una compañía de seguros para el ejercicio de su actividad. Efectivamente, toda empresa, a la hora de fijar los precios de los productos o servicios que ofrecen, tiene en mente sus costes internos, de forma que estos sean cubiertos.
- **Margen de beneficio**: una vez cubiertos los gastos, la compañía debe de aumentar un poco más el precio para poder obtener un beneficio y rentabilizar su actividad. El margen de beneficio de una aseguradora a otra podrá variar en función de la estrategia y el modelo de negocio que sigue.

El cálculo de una prima de un seguro general se basa en lo que se denomina como un **"equilibrio financiero-actuarial"** [8]. Esto quiere decir que,

---

[5] Se debe de aclarar que una póliza de seguro ofrece una serie de coberturas que varían en función de lo establecido en dicho contrato. Una cobertura representa el límite económico hasta el cual la compañía aseguradora está dispuesta a cubrir los daños producidos por un siniestro. De esta forma, las aseguradoras disponen de una cartera de pólizas muy amplia y heterogénea según las coberturas que ofrezcan cada una de ellas.

[6] Se consideran como gastos fijos aquellos que no varían durante el periodo en el que se desarrolla el ejercicio. Por ejemplo, el sueldo que se debe de pagar a los empleados.

[7] Los gastos variables son aquellos que varían en función del consumo (agua, electricidad, etc.).

para una póliza de seguro concreta, asumiendo que la función de utilidad asociada al tomador de la póliza es lineal, entonces se puede decir que la prima que deberá de pagar es proporcional al coste esperado de las pérdidas provocadas por el siniestro para el cual se ofrecen coberturas.

Asimismo, también se deben de cumplir una serie de **principios de tarificación** que buscan mantener un cierto nivel de precisión a la hora de evaluar los riesgos asociados a cierta póliza de seguro. La compañía aseguradora y los profesionales encargados de determinar el valor de las primas deben de escoger entre ellos el método que mejor se ajuste a sus necesidades. De entre estos principios se enumeran los siguientes principales:

- Principio de Equivalencia o de Prima Pura.
- Principio del Valor Esperado.
- Principio de la Varianza.
- Principio de la Desviación Típica.

El seguro trata de cubrir los riesgos asociados a la incertidumbre del futuro. Por tanto, se basa en la **probabilidad** de que un acontecimiento se produzca. De esta forma, tiene una base en las matemáticas estadísticas y probabilísticas. Y es que, a la hora de calcular el precio asociado a la prima de un seguro se debe de tener en cuenta esto. Es por ello por lo que se deben de definir los **modelos de frecuencia y de severidad** que van a ser empleados para el cálculo de la prima. Cuando se habla de los modelos de frecuencia se hace referencia a los modelos comúnmente conocidos y empleados como pueden ser la distribución de Poisson o la distribución Binomial. En cuanto a los modelos de severidad, se habla de otro tipo de distribuciones como la distribución Gamma o de Pareto.

Una vez definidos todos estos elementos se puede pasar a aplicar el cálculo de la prima de un seguro general. Este se hace, generalmente, mediante el método de la prima pura que define la tarifa de la siguiente manera, como se presenta en la **Ecuación 5**:

$$Tarifa = \frac{Prima\ Pura + Gastos\ Fijos\ por\ exposición}{1 - factor\ de\ Gastos\ Variables - factor\ de\ beneficio\ y\ contingencias}$$

Ecuación 5

2.1.3 Lenguaje de Programación

Para el desarrollo del trabajo se va a usar el lenguaje de programación Python. Se trata de un lenguaje que aparece hace 30 años, creado por Guido Van Rossum y cuyo nombre es un homenaje al grupo cómico británico Monty Python[8].

Este lenguaje de programación es "potente y fácil de aprender" [9].Además, es un lenguaje multiplataforma de código abierto, es decir, gratuito. Se emplea en aplicaciones web, en desarrollo software, la ciencia de datos o incluso en el Machine Learning. Se considera un **lenguaje de alto nivel orientado a objetos**.

Los **lenguajes** de programación **de alto nivel** se caracterizan por aproximarse bastante al lenguaje natural. El objetivo de estos es conservar la

[8] Formación teatral británica integrada por Graham Chapman, John Cleese, Terry Gilliam, Eric Idle, Terry Jones y Michael Palin.

independencia con el hardware, de modo que sea transparente para el programador. De este modo, son lenguajes que se pueden ejecutar en diferentes máquinas. Además, soportan distintos paradigmas tales como la programación orientada a objetos, funcional, estructurada, etc. Por su parte, es un lenguaje en el que los datos se estructuran de forma más compleja y dinámica con el uso de listas, pilas, mapas, etc. [10].

La **programación orientada a objetos** "se basa en el concepto de crear un modelo del problema de destino en sus programas. La programación orientada a objetos disminuye los errores y promociona la reutilización del código." [11].

Una de las áreas en las que se maneja este lenguaje es para la Inteligencia Artificial y el Machine Learning. Efectivamente, es uno de los lenguajes más populares en el mundo actualmente ya que su baja complejidad favorece el desarrollo de aplicaciones complejas. Además, dispone de una amplia selección de librerías para implementar soluciones basadas en Inteligencia Artificial. A continuación, se detallan algunas de ellas que pueden ser útiles para el desarrollo de este trabajo[12]:

- **Matplotlib**, para la visualización de los datos. Puede resultar útil a la hora de tratar los datos para comprobar algunos de los requisitos previos que impone el uso de algoritmos de Regresión Lineal como la normalidad de las series de datos.
- **Numpy**, como en el caso anterior, para visualizar datos y crear estructuras. Se trata de una librería que incorpora estructuras como arrays y matrices de grandes dimensiones. A diferencia de otras estructuras predefinidas de Python, las de esta librería se procesan mucho más rápido.
- **Pandas**, para el cálculo numérico y análisis de datos. Se trata de la librería de Python más usada por los científicos de datos hoy en día. Se emplea sobre todo para aplicaciones financieras y de ingeniería.
- **Scikit-learn**, para el entrenamiento del modelo. Efectivamente se suele usar para aprendizaje automático supervisado y no supervisado.
- **Seaborn**, para graficar los datos. Es una librería basada en Matplotlib que se usa, sobre todo, para graficar distribuciones aleatorias [13].
- **Statsmodels**, para analizar los resultados obtenidos. Efectivamente, completa la regresión lineal clásica, con nuevos estimadores de mínimos cuadrados [14].

2.1.4 Entorno de programación

Para programar el algoritmo usando el lenguaje Python, se necesita determinar el entorno en el que se van a ejecutar los scripts. En el caso de este trabajo se han decidido usar tres herramientas combinadas para formar el entorno.

Para empezar, se necesita un editor de texto y código fuente. Para ello, se usará **Visual Studio Code**. Se trata de una herramienta desarrollada por Microsoft compatible con los tres principales sistemas operativos actuales (Windows, macOS y Linux). La ventaja de usar este editor de texto es que dispone de una serie de extensiones, entre las que se encuentra aquella de Python, que permiten programar usando distintos lenguajes. No obstante, requiere que se descargue Python en el disco del ordenador con el que se está trabajando para poder instalar paquetes que no vienen predefinidos.

Para poder ejecutar desde Visual Studio Code los scripts de Python que se van a elaborar, se necesitan dos herramientas adicionales: Jupyter Notebook y Anaconda.

Jupyter Notebook es una interfaz web de código abierto que permite, entre otras cosas, la ejecución de código. Para ello, se basa en la comunicación con un núcleo (*kernel*, en inglés) que se debe de determinar. Aquí es donde interviene **Anaconda**. Se trata de una distribución libre de dos lenguajes de programación: Python y R, que busca dar apoyo a profesionales en los campos de la ciencia de datos y del aprendizaje automático. En este trabajo, se va a emplear para crear el núcleo desde el cual se van a ejecutar los scripts de Python.

2.1.5 Datos empleados

El cálculo de una prima de seguros de hogar necesita de una serie de datos acerca del parque de viviendas de una ciudad, como, por ejemplo, el año de construcción de la vivienda, los metros cuadrados habitables, etc. Asimismo, también se deben de tener en cuenta datos acerca de riesgos de incendio, terremotos, así como cualquier otro acontecimiento que pueda afectar a la integridad del hogar del asegurado.

Para ello, se va a hacer uso del **Portal de Datos Abiertos del Ayuntamiento de Madrid**. Se trata de un portal creado por el ayuntamiento de la ciudad de Madrid que proporciona datos abiertos, accesibles para el conocimiento de cualquier ciudadano. Esta base de datos recoge información estadística acerca de la ciudad de Madrid. Desde este portal, se puede acceder tanto a las estadísticas que mantiene el ayuntamiento acerca de la región, como a datos abiertos acerca de diferentes áreas. En este trabajo se va a hacer uso de los **datos estadísticos**. Estas estadísticas se encuentran organizadas en diferentes áreas de información, siendo de interés en este caso el área de **Edificación y Vivienda.** En esta área, se puede consultar la siguiente información:

- Planeamiento urbanístico.
- Censo de edificios y viviendas.
- Mercado de la vivienda.
- Estadística registral catastral (IBI).

De esta forma, se hará uso del **Censo de Edificios y Viviendas del año 2011** [15], ya que se trata de aquel más reciente. De estos datos estadísticos se extraerán las variables necesarias para realizar el cálculo de la prima de un seguro de hogar.

Asimismo, es necesario tener datos acerca de los importes de las primas de seguros para distintos tipos de pólizas. Como estos datos no están disponibles en bases de datos de acceso público, se tendrá que hacer uso de calculadoras de primas de seguros online, como, por ejemplo, **Rastreator.com**.

2.2 Motivación

Este trabajo viene motivado por las prácticas en empresa que realizó la autora. Efectivamente, la autora hizo unas prácticas en la empresa aseguradora Mapfre. Al no tener conocimientos previos acerca de este sector, pudo aprender los aspectos básicos.

Además, le pareció muy interesante cómo se calculan las primas de seguros, ya que se trata de un producto cuyo fin es proteger al asegurado de lo impredecible. Las predicciones que entran en juego a la hora de calcular el valor de una prima no son siempre de lo más acertadas, dependiendo del modelo estadístico y probabilístico usado.

Asimismo, la Inteligencia Artificial, con sus múltiples algoritmos de predicción permite resolver parte de esta problemática. Es por ello por lo que resultó interesante aplicar estas técnicas al cálculo de primas de seguros generales.

Por último, dado que los seguros generales abarcan un gran abanico de áreas, el trabajo se centra en los seguros de hogar dado la facilidad de acceso a los datos necesarios, así como al interés de la propia autora.

3 Metodología

En este apartado se presenta la metodología que se va a seguir para la elaboración del siguiente trabajo.

- **Estudio** y **comprensión** de la **tarificación en el seguro de hogar**.
- **Análisis sectorial**: ventajas sectoriales de la Inteligencia Artificial y presentación de tres casos de estudio que emplean métodos de Machine Learning para el seguro.
- **Consulta** y **recopilación** en un fichero Excel de datos de primas de seguros de hogar mediante un calculador de primas online.
- **Conversión** de la base de datos obtenida a un fichero csv y **carga** de los datos en el entorno de desarrollo.
- **Análisis** y **transformación** de los datos.
- **Construcción del modelo**: entrenamiento y predicción.
- **Evaluación del modelo** mediante el método del error cuadrático y del coeficiente de determinación: resultados generales y variables representativas.
- **Comparación** de los resultados obtenidos con los resultados reales.

4 Desarrollo

4.1 El seguro de hogar y su prima

El seguro de hogar se inscribe dentro de los seguros generales. Este, en particular, busca ofrecer coberturas a los **daños materiales sufridos dentro de una vivienda**[16] (humedades, incendios, robos, averías, etc.). Asimismo, este tipo de seguros también debe de cubrir aquellos desperfectos causados en las viviendas de terceros. Efectivamente, en las viviendas (sobre todo si se trata de viviendas adosadas o bloques de pisos) se pueden producir daños como humedades que pueden afectar a la vivienda de terceros. Es por ello por lo que los seguros de hogar, como otros seguros (automóvil, por ejemplo) tienen en cuenta también la **Responsabilidad Civil a terceros**.

Al calcular un seguro de hogar se deben de tener en cuenta dos elementos: el continente y el contenido. El **continente** hace referencia a la propia vivienda y todos los elementos que la componen (estructura, suministros, etc.). Por su parte, el **contenido** es todo aquello que se encuentra en el interior de la vivienda [17]. De esta forma, se deben de considerar los muebles, electrodomésticos, equipos informáticos, joyas, etc.

El continente es aquel que determinará de forma importante el precio de la prima que deberá de pagar el asegurado. De esta forma, los elementos concretos que influyen en este valor son los siguientes [18]–[20]:

- **Antigüedad** o Año de construcción de la vivienda: se contempla tanto la **fecha en la que se edificó** el inmueble, así como si, desde entonces, se han realizado **reformas**. A mayor antigüedad de la vivienda, existe un mayor riesgo de que surjan problemas, por lo que la prima se suele encarecer.
- **Materiales de construcción**: estos influyen en la **calidad** de la vivienda. Por una parte, si estos materiales son de una calidad alta, su coste de sustitución o reparación en caso de siniestro será mayor, por lo que afecta al alza el valor de la prima. Por otra parte, también se debe de hablar de la **robustez** de estos materiales. Se hace referencia a si se trata de materiales incombustibles, o combustibles hasta un cierto grado (hasta el 25% o más del 25%). De esta forma, si el hogar posee materiales incombustibles, el precio a pagar será menor que si son materiales más o menos combustibles (puesto que aumentan la probabilidad de sufrir un siniestro).
- **Tamaño**: cuando se habla del tamaño del hogar, se hace referencia tanto a los **metros cuadrados** de los que dispone, como al **número de habitaciones y baños o aseos**. Aquí el precio de la prima es proporcional al tamaño. Cuantos más metros y habitaciones, mayor será este.
- **Ubicación**: las condiciones de la zona, tanto **climatológicas** como de **seguridad** influyen en la determinación de la prima de la póliza.
- **Uso de la vivienda**: a las aseguradoras les resulta importante saber si el asegurado va a usar la vivienda como **residencia habitual** o como **residencia secundaria**. Se trata de datos cruciales en cuanto a estimar la Responsabilidad Civil que tiene el tomador de la póliza.
- **Altura del inmueble**: si se trata de un piso, se evalúa si se encuentra en un piso intermedio, planta baja, ático, primera planta.

Si se trata de una vivienda unifamiliar, si esta se encuentra aislada o adosada.

- **Sistemas de seguridad**: si la vivienda dispone de un sistema de alarma en todos los accesos, puerta blindada o acorazada o rejas en los accesos, el valor de la prima disminuye.

4.2 Aplicación de la Inteligencia Artificial en el sector asegurador: situación actual

La **Inteligencia Artificial** representa una tecnología en auge en la sociedad actual. La mediatización de los distintos avances realizados en este campo ha puesto de moda todas las herramientas basadas en ello. Y es que esta tecnología proporciona una serie de **ventajas** para cualquier sector de la economía. El sector asegurador no se queda al margen de estos avances y ya está empleando distintas técnicas basadas en Inteligencia Artificial para su desarrollo.

Efectivamente, se pueden encontrar numerosos **artículos académicos** sobre este tema enfocados, tanto en temas de tarificación de primas, como en temas de marketing y ventas.

En este apartado se quiere presentar una visión global acerca del estado del sector y de las distintas iniciativas elaboradas en base a esta tecnología.

4.2.1 Tarificación

La Inteligencia Artificial puede ser empleada a la hora de determinar el valor de la prima de un seguro. Efectivamente, se trata del mismo objetivo en torno al cual gira este trabajo.

Un ejemplo de ello se basa en un trabajo realizado en la **Universidad de Barcelona** [21] que presenta un caso de estudio centrado en la aplicación de distintos algoritmos de **Machine Learning** para la **tarificación de una prima para un seguro de automóvil** (que también es un tipo de seguro general, como se ha visto en el apartado introductorio de este trabajo).

En este trabajo, se busca evaluar la **precisión y robustez de una serie de algoritmos** de Machine Learning a la hora de calcular una tarifa. Estos algoritmos se clasifican dentro de los tres tipos existentes en la actualidad:

- **Aprendizaje supervisado**:
  - Regresión Lineal.
  - Regresión Logística.
  - Árbol de Decisión.
  - SVM (Support Vector Machine).
  - Naive Bayes.
- **Aprendizaje no supervisado**:
  - kNN (K-Nearest Neighbors).
  - Random Forest.
- **Aprendizaje reforzado**:
  - Generalized Boosted Regression Modelling.
  - Discrete Adaboost.

Los algoritmos se encuentran alimentados por una **cartera real de pólizas** de seguros de automóvil. Efectivamente, contienen datos procedentes de una muestra de jóvenes asegurados en el año 2011.

En este entorno, la variable dependiente corresponde a la existencia de una declaración de accidente con culpa. Las variables independientes son, entre otras, la edad del conductor asegurado, el sexo, los años que lleva con el carné de conducir, la antigüedad del vehículo, los kilómetros que recorre el tomador anualmente, etc.

Para construir los distintos modelos se ha decidido emplear el 75% de los datos como datos de entrenamiento, dejando el 25% restante para evaluar el modelo.

Asimismo, se utilizan dos métodos diferentes para el cálculo de primas: según el **coste esperado** y usando un **precio por kilómetro**.

De esta forma, se obtienen **conclusiones similares** para todos los modelos construidos. Efectivamente, los diferentes métodos **ofrecen capacidades de predicción de la siniestralidad muy similares**. No obstante, según el algoritmo empleado se observa una mayor o menor dispersión en los precios finales. El trabajo concluye que la elección del modelo dependerá del "deseo de interpretabilidad, reproducibilidad y elementos de la supervisión".

Así, en este caso, se emplea la Inteligencia Artificial para la tarificación de una prima. Para ello, se construyen distintos modelos que permiten predecir la siniestralidad (de un seguro de automóvil en este caso). Una vez obtenida esta, se calcula la prima como se ha visto en el apartado 2.1.2 de este trabajo.

4.2.2 Detección de fraudes

En el ámbito del seguro el **fraude** corresponde a todas aquellas acciones realizadas por los asegurados que buscan obtener cierto beneficio cuando no correspondería. También se puede hablar de fraude para hacer referencia a actos ilícitos llevados a cabo por las propias compañías de seguros [22]. Sea realizado por cualquiera de las partes implicadas en la póliza, el fraude representa un **delito tipificado en el Código Penal como estafa** [23].

Se trata de una realidad presente en el sector que representa un riesgo para los agentes económicos[9]. De esta forma, el **riesgo de fraude** es un parámetro que las aseguradoras deben de tener en cuenta a la hora de cuantificar la prima asociada a una póliza. Por tanto, cuanto menor sea este riesgo, menor será el impacto en el coste de la póliza. Es por ello por lo que, para las sociedades, la detección de estos fraudes resulta crucial.

En este contexto, se ha analizado cómo se puede aprovechar el desempeño de ciertos **algoritmos de Machine Learning para detectar fraudes**. Se va a presentar un trabajo centrado en la detección de fraudes en el seguro de automóvil realizado por catedráticos de la **Universitat de Valéncia** [24].

En él se busca comparar el desempeño de una serie de algoritmos de Machine Learning para la **predicción de siniestros fraudulentos**. Para el estudio se emplean datos extraídos de una base de datos de **pólizas de automóviles de una entidad aseguradora** que tiene presencia en el ramo de la automoción. Cabe destacar que los siniestros fraudulentos suelen representar un porcentaje reducido sobre el total de siniestros. Es por ello por lo que, de la

[9] Se definen como agentes económicos a todas las partes implicadas en el contrato de seguro que pueden incurrir en una actividad fraudulenta.

base de datos, se obtiene una muestra lo suficientemente significativa donde un 56% de los siniestros son legítimos y el resto son fraudulentos.

Los diferentes algoritmos de Machine Learning que se van a emplear para el estudio son los siguientes:

- Regresión Logística (Logit).
- Máquinas de Vector de Soporte (SVM).
- Árboles de Decisión (DT).
- Naïve Bayes (NB).
- Redes Neuronales (NNET).
- Bosques Aleatorios (RF).
- Extreme Gradient Boosting (XGB).

La comparación de los distintos modelos se ha realizado en base a distintas medidas. De esta forma, el estudio concluye que, en cuanto a la **idoneidad** del modelo no se encuentra uno mejor que otro, puesto que obtienen una **precisión muy similar**. No obstante, si se habla de **capacidad predictiva** se ha visto que el modelo basado en **Regresión Logística** tiene un mejor desempeño.

De esta forma, a través de este trabajo, se puede ver cómo la Inteligencia Artificial puede resultar de utilidad para predecir futuros siniestros fraudulentos y, por tanto, prevenirlos. Esto permitiría a las aseguradoras mitigar su riesgo de fraude, reduciendo el impacto que este tiene sobre las primas que pagan sus clientes.

4.2.3 Segmentación de clientes

Otro ámbito en el que el uso de técnicas basadas en la Inteligencia Artificial puede ser beneficios es aquel de **Marketing y Ventas**. Efectivamente, estas metodologías permiten trabajar con una mayor cantidad de datos, así como identificar patrones. Esto último puede ayudar a **agrupar clientes**, **identificar comportamientos de compra**, etc., que pueden ser muy útiles a la hora de elaborar las **estrategias de ventas**. Se trata de conocer mejor al cliente para poder identificar mejor sus necesidades y, por tanto, **personalizar** los productos que se le ofrecen. Esto genera, por lo general, una mayor satisfacción que puede llegar a fidelizar al cliente. Para toda organización con ánimo de lucro esto resulta fundamental para garantizar una fuente de ingresos estable.

Por tanto, en este apartado se estudia un trabajo publicado por la **Escuela de Administración de Negocios para Graduados** (ESAN, situada en Lima, Perú) [25]. El trabajo se centra en la clasificación de clientes en una empresa de seguros mediante técnicas de Machine Learning. Se busca elaborar un **modelo predictivo destinado a identificar la póliza que mejor se ajuste a un cliente**, de forma que se le pueda ofrecer.

En este caso, como en los apartados anteriores, se pone el foco en el ramo del automóvil. Los datos empleados provienen de una empresa aseguradora peruana y son del año 2019. El trabajo se descompone en las siguientes fases o etapas:

- Recopilación de la base de datos.
- Limpieza y preprocesamiento de datos.
- Desarrollo del modelo de clasificación.

Para la clasificación de los clientes se van a emplear y comparar dos algoritmos de Machine Learning: k-NN (K-Nearest Neighbors) y Regresión Logística.

Las conclusiones evidencian que el modelo que ofrece una **mayor precisión** es aquel basado en el algoritmo **k-NN**.

4.3 Fase 1: Configuración del entorno de programación

Antes de poder empezar a desarrollar el código e implementar y probar el algoritmo, es necesario determinar el entorno en el que se va a trabajar. En este caso, tan sólo se necesita crear el núcleo en el que se van a ejecutar los scripts.

De esta forma, es necesario abrir un terminal de Anaconda. Desde el terminal, se ejecuta el comando: ***conda create -n myenv python=3.10 pandas jupyter seaborn scikit-learn keras tensorflow***, para crear el núcleo que se llamará *myenv*. Una vez creado, desde Visual Studio Code, usando en editor de texto de Jupyter se selecciona *myenv* como el núcleo.

4.4 Fase 2: Modelado de datos

Los datos empleados para nutrir el algoritmo, como se ha explicado anteriormente, provienen de dos fuentes distintas. Se ha recopilado un total de 175.392 datos organizados en 9 columnas como lo ilustra la **Figura 1**.

Distrito	Código Postal	Año de construcción	Tipo de vivienda	Materiales	Metros	Nº habitaciones	Nº aseos	Prima (€/año)
01. Centro	28013	1900	Piso en planta intermedia	Incombustibles	25	1	1	198,67
01. Centro	28013	1900	Piso en planta baja	Incombustibles	25	1	1	223,67
01. Centro	28013	1900	Piso en primera planta	Incombustibles	25	1	1	201,33
01. Centro	28013	1900	Ático o último piso	Incombustibles	25	1	1	165,00
01. Centro	28013	1900	Piso en planta intermedia	Combustibles hasta 25%	25	1	1	198,00
01. Centro	28013	1900	Piso en planta baja	Combustibles hasta 25%	25	1	1	217,50
01. Centro	28013	1900	Piso en primera planta	Combustibles hasta 25%	25	1	1	198,00
01. Centro	28013	1900	Ático o último piso	Combustibles hasta 25%	25	1	1	215,00
01. Centro	28013	1900	Piso en planta intermedia	Combustibles más 25%	25	1	1	200,00
01. Centro	28013	1900	Piso en planta baja	Combustibles más 25%	25	1	1	219,50
01. Centro	28013	1900	Piso en primera planta	Combustibles más 25%	25	1	1	200,00
01. Centro	28013	1900	Ático o último piso	Combustibles más 25%	25	1	1	217,50
01. Centro	28013	1900	Piso en planta intermedia	Incombustibles	35	2	1	207,00
01. Centro	28013	1900	Piso en planta baja	Incombustibles	35	2	1	234,00
01. Centro	28013	1900	Piso en primera planta	Incombustibles	35	2	1	128,33
01. Centro	28013	1900	Ático o último piso	Incombustibles	35	2	1	202,00
01. Centro	28013	1900	Piso en planta intermedia	Combustibles hasta 25%	35	2	1	207,00
01. Centro	28013	1900	Piso en planta baja	Combustibles hasta 25%	35	2	1	291,00
01. Centro	28013	1900	Piso en primera planta	Combustibles hasta 25%	35	2	1	207,00
01. Centro	28013	1900	Ático o último piso	Combustibles hasta 25%	35	2	1	291,00
01. Centro	28013	1900	Piso en planta intermedia	Combustibles más 25%	35	2	1	209,00
01. Centro	28013	1900	Piso en planta baja	Combustibles más 25%	35	2	1	231,00
01. Centro	28013	1900	Piso en primera planta	Combustibles más 25%	35	2	1	209,00
01. Centro	28013	1900	Ático o último piso	Combustibles más 25%	35	2	1	228,00
01. Centro	28013	1900	Piso en planta intermedia	Incombustibles	58	3	1	219,67
01. Centro	28013	1900	Piso en planta baja	Incombustibles	58	3	1	248,33
01. Centro	28013	1900	Piso en primera planta	Incombustibles	58	3	1	222,33
01. Centro	28013	1900	Ático o último piso	Incombustibles	58	3	1	243,33
01. Centro	28013	1900	Piso en planta intermedia	Combustibles hasta 25%	58	3	1	236,50
01. Centro	28013	1900	Piso en planta baja	Combustibles hasta 25%	58	3	1	262,50
01. Centro	28013	1900	Piso en primera planta	Combustibles hasta 25%	58	3	1	273,00
01. Centro	28013	1900	Ático o último piso	Combustibles hasta 25%	58	3	1	258,00

Figura 1: Base de datos acerca de las primas de un seguro de hogar creada para el desarrollo del proyecto, Fuente: Elaboración propia.

Estos se encuentran almacenados en un archivo **.csv** que contiene los siguientes campos:

- **Distrito**: se trata de un campo de datos de tipo texto.

- **Código postal**: se trata de un campo de datos de tipo número entero.
- **Año de construcción**: se trata de un campo de datos de tipo número entero.
- **Tipo de vivienda**: se trata de un campo de datos de tipo texto.
- **Materiales**: se trata de un campo de datos de tipo texto.
- **Metros**: se trata de un campo de datos de tipo número entero.
- **Número habitaciones**: se trata de un campo de datos de tipo texto. Efectivamente, esta columna contiene caracteres no numéricos (por ejemplo: 6+), por lo que no se reconoce con un tipo de dato numérico.
- **Número aseos**: se trata de un campo de datos de tipo texto. Aquí sucede lo mismo que en el caso del número de habitaciones.
- **Prima (€/año)**: se trata de un campo de datos de tipo número decimal.

Para implementar el algoritmo y empezar con la fase de entrenamiento de este, se deben de proveer valores numéricos. Efectivamente, los algoritmos de regresión lineal se usan, típicamente, para **valores continuos**. Esto hace referencia a aquellos valores que se pueden medir de forma continua en una escala (por ejemplo, la altura, el tiempo, etc.). No obstante, también se pueden emplear variables de tipo **categóricas**. Se trata de variables que permiten clasificar elementos, por lo que adoptan un formato de texto.

En este trabajo se están manejando tanto variables continuas como variables categóricas. Al cargar el fichero, se puede consultar el tipo de datos que tiene cada una de las variables enumeradas anteriormente gracias a la sentencia ***data.dtypes***. De esta forma, se puede observar esto en la **Figura 2**:

```
Distrito              object
Código Postal          int64
Año de construcción    int64
Tipo de vivienda      object
Materiales            object
Metros                 int64
Nº habitaciones       object
Nº aseos              object
Prima (€/año)         object
dtype: object
```

Figura 2: Tipo de datos inicial de cada una de las variables del dataset, Fuente: Elaboración propia.

Se observa que los únicos campos reconocidos como variables numéricas son “Código Postal”, “Año de construcción” y “Metros”. No obstante, los campos “Nº habitaciones”, “Nº aseos” y “Prima (€/año)”, también son campos que contienen valores numéricos, pero son interpretados por la máquina como texto.

Por una parte, en el caso del campo “Prima (€/año)”, el problema reside en que se trata de un campo de números decimales. En el fichero de origen, los decimales vienen delimitados por una coma y no por un punto, dando pie a que se interprete erróneamente. Por tanto, es necesario reemplazar las comas por puntos usando ***str.replace()***. De esta forma, se puede aplicar la conversión a números decimales al campo.

Por otra parte, para los campos que recogen el número de habitaciones y de baños, el primer paso que se debe de llevar a cabo consiste en cambiar ambos campos para que contengan exclusivamente valores numéricos.

El campo “Nº habitaciones” puede tomar los siguientes valores:

- 1
- 2
- 3
- 4
- 5
- 6+

Por tanto, contiene valores numéricos excepto en el último caso que, al llevar el carácter “+”, hace que el algoritmo los interprete como texto. Si bien es cierto que existen varios métodos que permiten pasar texto a números, en este caso se puede cambiar este campo por el valor 6, eliminando el carácter de la suma. Efectivamente, el campo representa la opción de que la vivienda tenga 6 o más habitaciones, por lo que parece razonable realizar este cambio manualmente.

El campo “Nº aseos” puede tomar los siguientes valores:

- 1
- 2
- 3
- 3+

Como en el caso anterior, el último valor hace que se interprete el campo como un campo de texto. En este caso, parece razonable cambiar el valor “3+” por un “4”, ya que el valor “3” ya está presente.

Una vez realizados estos cambios, los tipos de datos de los diferentes campos que componen el conjunto de datos con el que se va a trabajar quedan como se puede ver en la **Figura 3**:

```
Distrito               object
Código Postal           int64
Año de construcción     int64
Tipo de vivienda       object
Materiales             object
Metros                  int64
Nº habitaciones         int32
Nº aseos                int32
Prima (€/año)         float64
dtype: object
```

Figura 3: Tipo de datos de las distintas variables del dataset tras las transformaciones, Fuente: Elaboración propia.

De esta forma, tan sólo quedan tres campos que son variables categóricas: “Distrito”, “Tipo de vivienda” y “Materiales”. Para poder trabajar con estas variables con el algoritmo de Regresión Lineal, se deben de convertir en valores continuos.

Para ello, se plantean dos posibles opciones. Por una parte, se puede emplear el método de **Label Encoding**. Se trata de un enfoque que busca etiquetar las variables categóricas con números, normalmente, enteros. Para

ello, se identifican los diferentes valores que puede tomar la variable en cuestión y se asigna, a cada uno de ellos, un número, empezando por el número 0. Así, si se dispone, por ejemplo, de una variable Dificultad que puede tomar los valores: Baja, Media y Alta, este método los clasificaría del siguiente modo:

- Baja: 0.
- Media: 1.
- Alta: 2.

Resulta un método eficaz e intuitivo, no obstante, para ciertos algoritmos de aprendizaje automático, como el que se emplea en este trabajo, este enfoque puede provocar errores. Para explicarlo, se debe de recordar la función que define el modelo de Regresión Lineal es aquella definida por la **Ecuación 2**.

X representa la variable dependiente que, en este caso, es la variable categórica convertida a una variable continua. Cuando esta toma el valor Alta, entonces X= 2. Por tanto, la ecuación queda de la siguiente manera, como se puede ver en la **Ecuación 6**:

$$Y = 2\beta 0 + \beta 1 + \varepsilon$$

Ecuación 6

De este modo, se está asumiendo que la diferencia entre la dificultad baja, que es el primer grupo y la dificultad alta, tercer grupo, es dos veces mayor a aquella entre la dificultad baja y media. Esto podría ser cierto en algunos casos, pero no en todos. Por tanto, se debe de emplear otro enfoque para convertir las variables de tipo texto a números, para evitar posibles errores.

Existen varias formas de hacer esto. Por una parte, se pueden crear las llamadas **variables *dummies***. Son variables ficticias que suelen tomar valores binarios. De esta forma, el conjunto de datos con el que se va a trabajar en la siguiente fase sería aquel ilustrado por la **Figura 4**:

	Código Postal	Año de construcción	Metros	Nº habitaciones	Nº aseos	Prima (€/año)	Distrito_02. Arganzuela	Distrito_03. Retiro	Distrito_04. Salamanca	Distrito_05. Chamartín	...	Distrito_17. Villaverde	Distrito_18. Villa Vallecas	Distrito_19. Vicálvaro	Distrito_20. San Blas-Canillejas	Distrito_21. Barajas	vi
0	28013	1900	25	1	1	198.67	0	0	0	0	...	0	0	0	0	0	
1	28013	1900	25	1	1	223.67	0	0	0	0	...	0	0	0	0	0	
2	28013	1900	25	1	1	201.33	0	0	0	0	...	0	0	0	0	0	
3	28013	1900	25	1	1	165.00	0	0	0	0	...	0	0	0	0	0	
4	28013	1900	25	1	1	198.00	0	0	0	0	...	0	0	0	0	0	

Figura 4: Conjunto de datos con variables dummies, Fuente: Elaboración propia.

Por otra parte, existe otra metodología conocida como **One Hot Encoding**. Se busca crear una columna binaria para cada valor que pueda tomar la variable categórica. Cuando, en un registro, el valor está presente, se marca la columna con un 1. De lo contrario, aparece codificada con un 0, como en el caso de las variables dummies. Sin embargo, con este método, cada registro del conjunto de datos contiene un vector binario que funciona como un indicador acerca de la presencia de la variable categórica.

Para realizar este cambio, es necesario seleccionar las variables categóricas por separado. Las columnas que contienen las variables categóricas se guardan en una variable distinta. Una vez hecho esto, se puede aplicar el método. Para ello, primero se necesita crear una instancia de **One Hot Encoder**. Python ofrece la función **OneHotEncoder()** de la librería skicit-learn. A la

instancia creada se le aplica la función **fit_transform()** a las variables categóricas, lo que devuelve un array con siete columnas rellenadas con ceros y unos. Estas columnas no tienen nombres, sino que están etiquetadas con números del 0 al 6, en este caso.

Habiendo realizado la transformación de las variables categóricas a valores numéricos, ya que se quiere trabajar con el conjunto del dataframe inicial, se deben de añadir al conjunto de datos inicial las nuevas columnas creadas empleando la función **join()**, y eliminando las columnas que contienen datos categóricos con la función **drop()**. También se pueden renombrar las columnas para una mejor comprensión. De esta forma, el conjunto de datos final sería el expuesto en la **Figura 5**:

Año d...	Metros	Nº ha...	Nº ase...	Planta...	Planta...	Prime...	Ático	Comb...	Comb...	Incom...
1900	25	1	1	0	1	0	0	0	0	1
1900	25	1	1	1	0	0	0	0	0	1
1900	25	1	1	0	0	1	0	0	0	1
1900	25	1	1	0	0	0	1	0	0	1
1900	25	1	1	0	1	0	0	1	0	0
1900	25	1	1	1	0	0	0	1	0	0
1900	25	1	1	0	0	1	0	1	0	0
1900	25	1	1	0	0	0	1	1	0	0
1900	25	1	1	0	1	0	0	0	1	0
1900	25	1	1	1	0	0	0	0	1	0
1900	25	1	1	0	0	1	0	0	1	0
1900	25	1	1	0	0	0	1	0	1	0

Figura 5: Conjunto de datos con variables categóricas transformadas mediante el método One Hot Encoder, Fuente: Elaboración propia.

Al tener estos dos métodos que proporcionan resultados distintos, para el resto del desarrollo de este trabajo, se va a trabajar con ambos. De esta forma, se pueden comparar los resultados finales obtenidos y determinar en qué medida afecta el método empleado al desempeño del algoritmo o si, por lo contrario, el método usado no tiene ningún efecto.

4.5 Fase 3: Comprobación de linealidad

Antes de poder pasar al entrenamiento del modelo es importante comprobar la hipótesis principal en la que se basa el algoritmo de Regresión Lineal: la linealidad entre las variables independientes y aquella dependiente.

Para ello, se puede representar la relación mediante gráficos de dispersión ya que muestran bien la tendencia que sigue una distribución de datos. Haciendo esta representación, se puede observar en los **Gráficos 1 a 4**, que las variables parecen mantener cierta relación lineal.

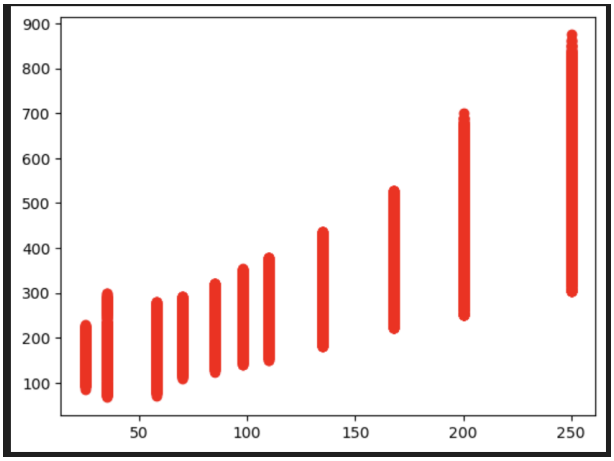


Gráfico 1: Prima de un seguro de hogar en función de sus metros cuadrados, Fuente: Elaboración propia.

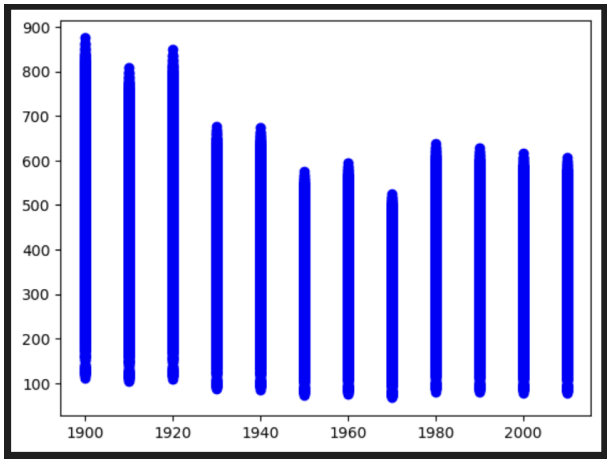


Gráfico 2: Prima de un seguro de hogar en función de su año de construcción, Fuente: Elaboración propia.

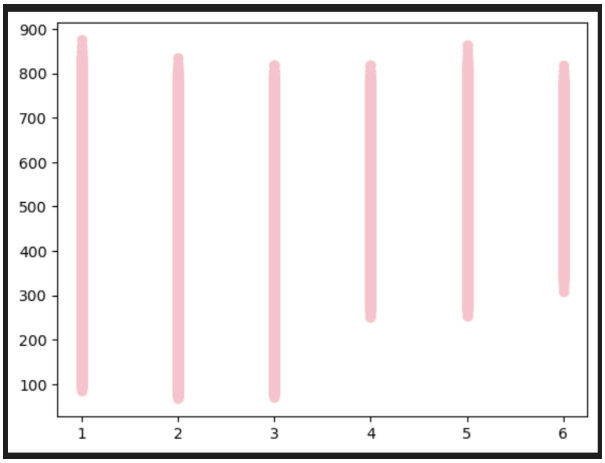


Gráfico 3: Prima de un seguro de hogar en función del número de habitaciones, Fuente: Elaboración propia.

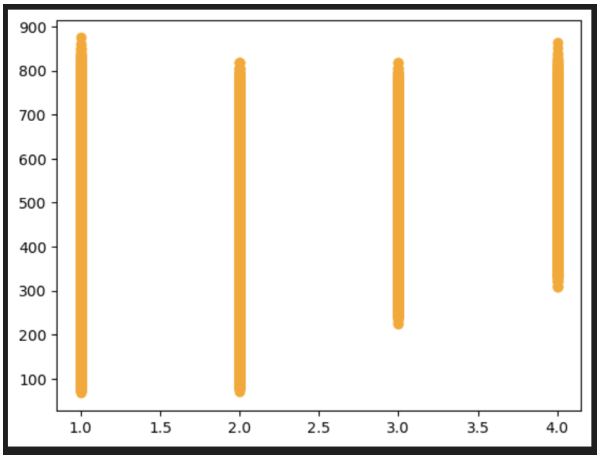


Gráfico 4: Prima de un seguro de hogar en función del número de baños/aseos, Fuente: Elaboración propia.

Se pueden observar las siguientes relaciones:

- El **Gráfico 1** muestra que cuantos más metros tenga la vivienda, mayor será la prima que deberá de pagar el tomador de la póliza.
- El **Gráfico 2** ilustra que cuanto más recientemente se haya construido la vivienda, menor será la prima ofrecida en la póliza.
- El **Gráfico 3** permite ver que, a mayor número de habitaciones, mayor será la prima.
- El **Gráfico 4** muestra esto mismo extrapolado al número de baños/aseos de los que disponga un hogar.

No obstante, en el caso del Código Postal, ilustrado en el **Gráfico 5**, no parece haber relación alguna. Por tanto, puede resultar razonable eliminar esta variable del conjunto de datos de cara a las siguientes fases.

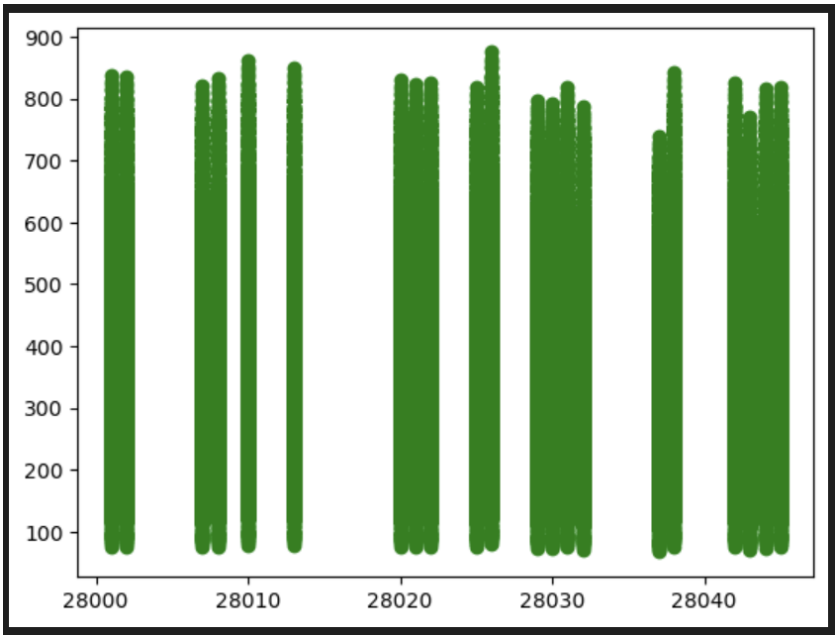


Gráfico 5: Prima de un seguro de hogar en función de su código postal, Fuente: Elaboración propia.

4.6 Fase 4: División del conjunto de datos y entrenamiento del modelo

Una vez modelados los datos, se va a pasar a entrenar el modelo. Cabe destacar que, en este trabajo, se está trabajando con más de dos variables. Por tanto, el modelo que se va a desarrollar consiste en un **modelo de Regresión Lineal Múltiple**. Sabiendo el tipo de Regresión Lineal que se va a emplear se puede empezar la fase de entrenamiento.

Antes de empezar es necesario dividir el dataframe distinguiendo las variables independientes de aquella dependiente. En este caso, los dos subconjuntos de datos resultantes serían los siguientes:

- **Variables independientes (X)**: Año de construcción, Tipo de vivienda, Materiales, Metros, Nº habitaciones, Nº aseos. En este caso se han descartado las variables Distrito y Código Postal ya que no cumplían la hipótesis de linealidad. Cabe destacar que, aunque se haya evaluado la linealidad de la segunda variable, al estar ligada con la primera, se pueden eliminar ambas, pues se considera que se obtendría el mismo resultado.
- **Variable dependiente (y)**: Prima (€/año).

Tras haber hecho esta separación, se dividen ambos subconjuntos de nuevo. Por una parte, se tendrá el conjunto de datos de entrenamiento y, por otra, los datos que serán empleados para probar el modelo.

- **Datos de entrenamiento**: se trata del conjunto de datos del dataframe que serán empleados para entrenar el modelo. De esta forma, a partir de ellos, el algoritmo podrá encontrar relaciones y determinar las ponderaciones correspondientes a cada una de las variables independientes.

- **Conjunto de prueba:** se trata de aquellos datos con los que se va a probar el modelo. Normalmente, el tamaño del conjunto de prueba suele ser, como máximo, el 25% de los datos empleados.

Para entrenar el modelo se emplea la función de la librería skicit-learn **train_test_split()**. Esta función recibe una serie de parámetros.

- ***arrays**: se trata de listas de datos que deben de tener el mismo tamaño/dimensión, sobre los cuales se quiere aplicar el modelo. En el caso de este trabajo, se tendrán que pasar dos listas. Por una parte, se pasará la lista **X** que contiene los datos de las variables independientes. Por otra parte, también se debe proveer la lista **y** que contiene los datos de la variable dependiente.
- **test_size**: es una variable numérica que puede ser entera o decimal. Representa la proporción o el número entero de datos sobre los que se quieren realizar las pruebas. En el marco de este trabajo se ha decidido que resulta razonable determinar un conjunto de prueba del 20% de los datos originales. De esta forma se inicializa a **0.2**[10].
- **random_state**: es un parámetro de tipo entero que controla el generador de número aleatorios. Es lo que permite reproducir la función un cierto número de veces, obteniendo para cada valor un conjunto de datos único. En este caso, se empieza con un valor de **42**. Aunque puede resultar interesante probar otros valores y ver cómo varía la precisión del modelo.

Una vez divididos los datos, se usa el conjunto de entrenamiento para entrenar el modelo. Para este entrenamiento es necesario, antes de nada, especificar el modelo que se quiere emplear. En el marco de este trabajo se deberá de especificar que se quiere trabajar con Regresión Lineal. Tras esto, se usa la función **fit()**[11].

4.7 Fase 5: Predicción y Evaluación del modelo

Esta fase representa la última fase en la construcción del modelo de Regresión Lineal. Habiendo entrenado el modelo, se pretende evaluar la calidad del entrenamiento realizando predicciones acerca de la variable dependiente que, en el ámbito de este trabajo es la prima anual de un seguro de hogar.

Estas predicciones se basan en las relaciones y los pesos que el algoritmo ha identificado y asignado a cada una de las variables independientes en la fase anterior. Para realizarlas usamos la función **predict()** que recibe como parámetros el conjunto de datos de prueba definido anteriormente.

Existen varios métodos que permiten evaluar el modelo. En este trabajo usaremos dos medidas distintas para obtener un análisis más exhaustivo.

---

[10] Esto significa que el 80% de los datos será empleado para entrenar el modelo.
[11] Se encarga de ajustar los parámetros de Regresión Lineal a los datos empleados. De esta forma, se busca asignar un peso a cada uno de los coeficientes de las variables independientes x_1, …, x_n para obtener el resultado esperado en la variable dependiente y.

Uno de los métodos más empleados se basa en el cálculo del **error cuadrático**. El error cuadrático (Mean Squared Error – MSE, en inglés) mide el **promedio de la diferencia entre los valores reales y los valores estimados**, es decir, los **errores**, elevados al cuadrado. Se trata de un método comúnmente empleado a la hora de evaluar modelos de regresión. Cuanto más pequeño sea el valor obtenido, mejor desempeño tendrá el modelo, puesto que significa que la diferencia entre los valores estimados y los reales es menor.

Otro método que puede resultar útil a la hora de evaluar la precisión del modelo consiste en calcular el **coeficiente de determinación** o **R-cuadrado**. Es una medida estadística que indica el **porcentaje de varianza en la variable dependiente que se puede explicar por las variables independientes**. Puede tomar valores entre el 0% y el 100%. A diferencia del error cuadrático, cuanto más alto sea el valor obtenido, es decir, cuanto más cercano al 100% sea, mejor será el modelo.

Por último, la forma más rudimentaria de evaluar el modelo sería graficar los resultados obtenidos frente a los esperados. Trazando conjuntamente la línea de regresión se puede observar cómo de bien se ajustan los datos a esta.

5 Análisis comparativo de los resultados obtenidos

Respecto de los valores reales, obtenidos de calculadores de primas de seguros online, se ha obtenido un error cuadrático relativamente alto, aunque el coeficiente de determinación se sitúa en un 89% de precisión, lo que representa un valor elevado en este aspecto.

5.1 Resultados obtenidos según método de transformación de variables categóricas

En este apartado se pretende discernir si el método empleado para transformar las variables categóricas presentes en el conjunto de datos inicial juega un papel importante en el modelo.

Cabe destacar que se trata de dos métodos muy similares, por lo que se podría esperar que el impacto fuera mínimo. Esto se puede ver comparando los resultados obtenidos.

Por un lado, la transformación mediante **variables dummies** presenta los resultados ilustrados en la **Figura 6**:

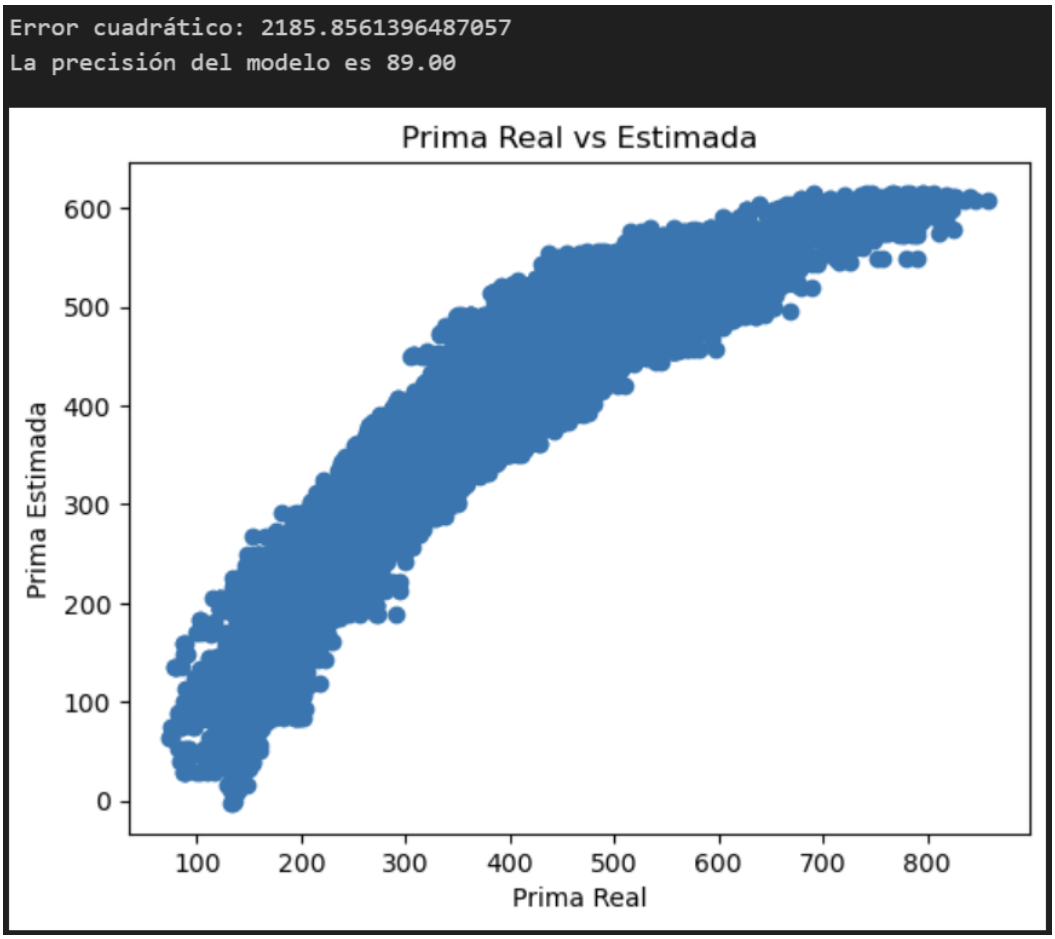


Figura 6: Error cuadrático, coeficiente de determinación y representación gráfica de las primas predichos por el modelo frente a las primas reales usando variables dummies, Fuente: Elaboración propia.

Se ha obtenido un **error cuadrático de 2185.86**. Se podría considerar como un valor elevado. No obstante, no queda del todo claro qué es un valor pequeño en este caso. Es por ello por lo que el coeficiente de determinación nos permite aportar mayor claridad acerca de la precisión del modelo. En este caso, se ha obtenido una **precisión del 89%**. Se trata de un valor muy cercano al 90%, lo que podría indicar que el modelo construido podría predecir con cierta exactitud la prima de un seguro de hogar.

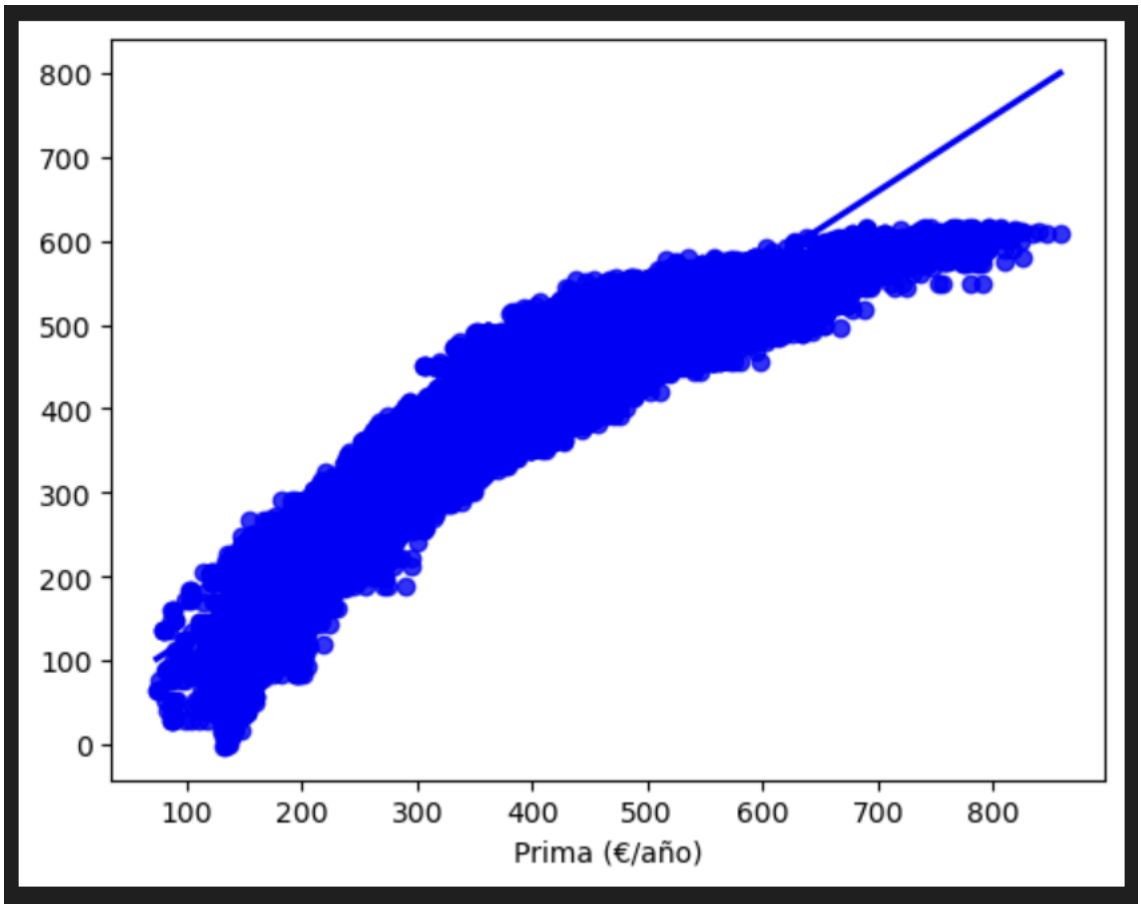


Figura 7: Representación gráfica de las primas predichas por el modelo en función de las primas reales junto con su recta de regresión, Fuente: Elaboración propia.

Por último, trazando la línea de regresión se puede apreciar en la **Figura 7** un mayor ajuste de los datos para primas entre 0 €/año y 600 €/año. A partir de este valor las predicciones obtenidas se encuentran muy alejadas de la línea trazada. Por tanto, se podría decir que el modelo puede resultar útil para primas hasta 600 €/año, pero no para predecir valores de primas cuyo importe es mayor.

Por otro lado, la transformación de las variables categóricas mediante el método **One Hot Encoder** presenta los resultados expuestos en la **Figura 8**:

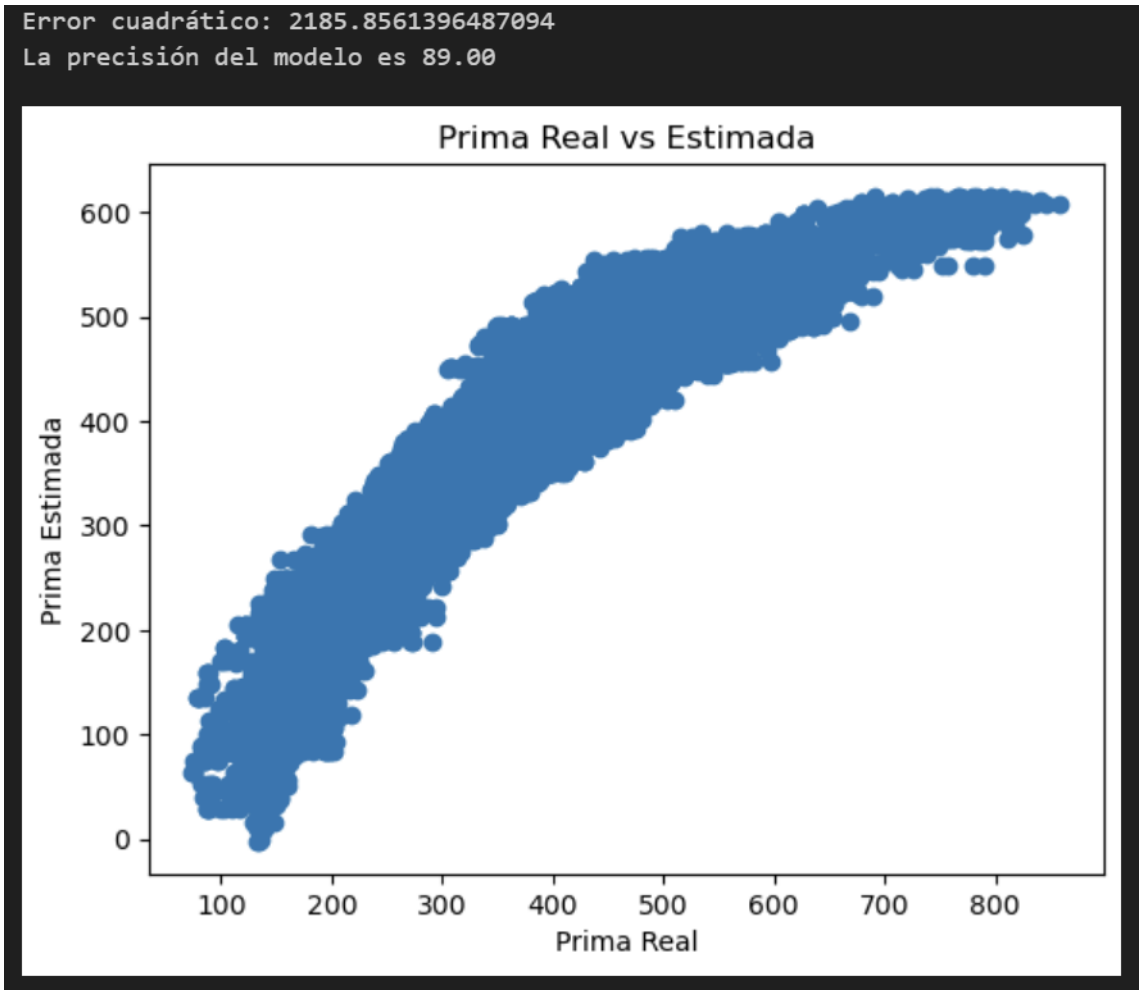


Figura 8: Error cuadrático, coeficiente de determinación y representación gráfica de las primas predichas por el modelo frente a las primas reales usando One Hot Encoder, Fuente: Elaboración propia.

Si se compara con el caso anterior se puede ver que se obtienen prácticamente los mismos resultados. La única diferencia reside en los decimales obtenidos en el error cuadrático. No obstante, al tratarse de valores tan pequeños se puede considerar esta como despreciable. Graficando los datos junto con la línea de regresión obtenemos la misma figura, por lo que las conclusiones expuestas anteriormente siguen siendo válidas.

De esta forma, se ha visto que, en este caso, como se esperaba, la diferencia entre el método de transformación empleado es prácticamente inexistente. Por tanto, en el modelo construido, ambos métodos son válidos.

5.2 Resultados obtenidos según separación de datos de entrenamiento y conjunto de prueba

En este apartado se quiere ver si variaciones en el tamaño y distribución de los conjuntos de entrenamiento y prueba tienen un impacto significativo en la precisión del modelo. Los resultados obtenidos en el apartado anterior servirán como base para realizar esta comparación.

Para empezar, se puede analizar cómo el tamaño del conjunto de datos de entrenamiento afecta al modelo. El tamaño inicial escogido es del 20%. Por tanto, se quiere analizar el resultado obtenido con los siguientes valores:

- **Tamaño del 25%:** se trata del tamaño máximo recomendado para los datos de entrenamiento. En este caso, en la **Figura 9**, se pueden observar dos tendencias. Por una parte, respecto de los valores iniciales, el error cuadrático aumenta ligeramente, por lo que indicaría que la precisión del modelo sería menor. No obstante, por otra parte, el coeficiente de determinación también aumenta ligeramente, indicando una mejora en la precisión del modelo. Aunque cabe destacar que se trata de variaciones muy pequeñas en los valores y, por tanto, poco significativas.

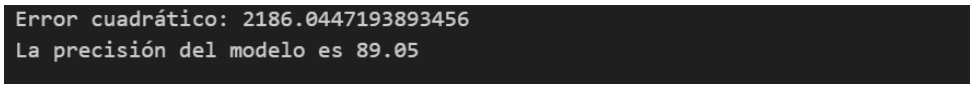


Figura 9: Error cuadrático y coeficiente de determinación para un tamaño del conjunto de entrenamiento del 25%, Fuente: Elaboración propia.

- **Tamaño del 15%:** en este caso, en la **Figura 10**, se observa un aumento en el valor del error cuadrático mucho mayor respecto del valor inicial. Por tanto, sí que se tendría una menor precisión en este modelo. Asimismo, el valor obtenido en el coeficiente de determinación corrobora esto último.

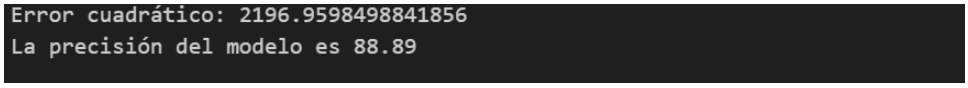


Figura 10: Error cuadrático y coeficiente de determinación para un tamaño del conjunto de entrenamiento del 15%, Fuente: Elaboración propia.

- **Tamaño del 10%:** respecto del caso anterior se observa un aumento del error cuadrático y una reducción del coeficiente de determinación en la **Figura 11**, reduciendo la precisión del modelo.

Figura 11: Error cuadrático y coeficiente de determinación para un tamaño del conjunto de entrenamiento del 10%, Fuente: Elaboración propia.

- **Tamaño del 5%:** resulta interesante notar que, en la **Figura 12** el error cuadrático resulta menor que para los tamaños del 15% y 10%, aunque el coeficiente de determinación es menor. También se está perdiendo precisión en este caso.

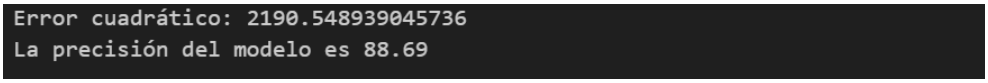


Figura 12: Error cuadrático y coeficiente de determinación para un tamaño del conjunto de entrenamiento del 5%, Fuente: Elaboración propia.

Por tanto, el valor asignado inicial parece ser el valor óptimo puesto que aporta la mayor precisión. En cuanto al valor de la variable **random_state**, se puede observar lo siguiente:

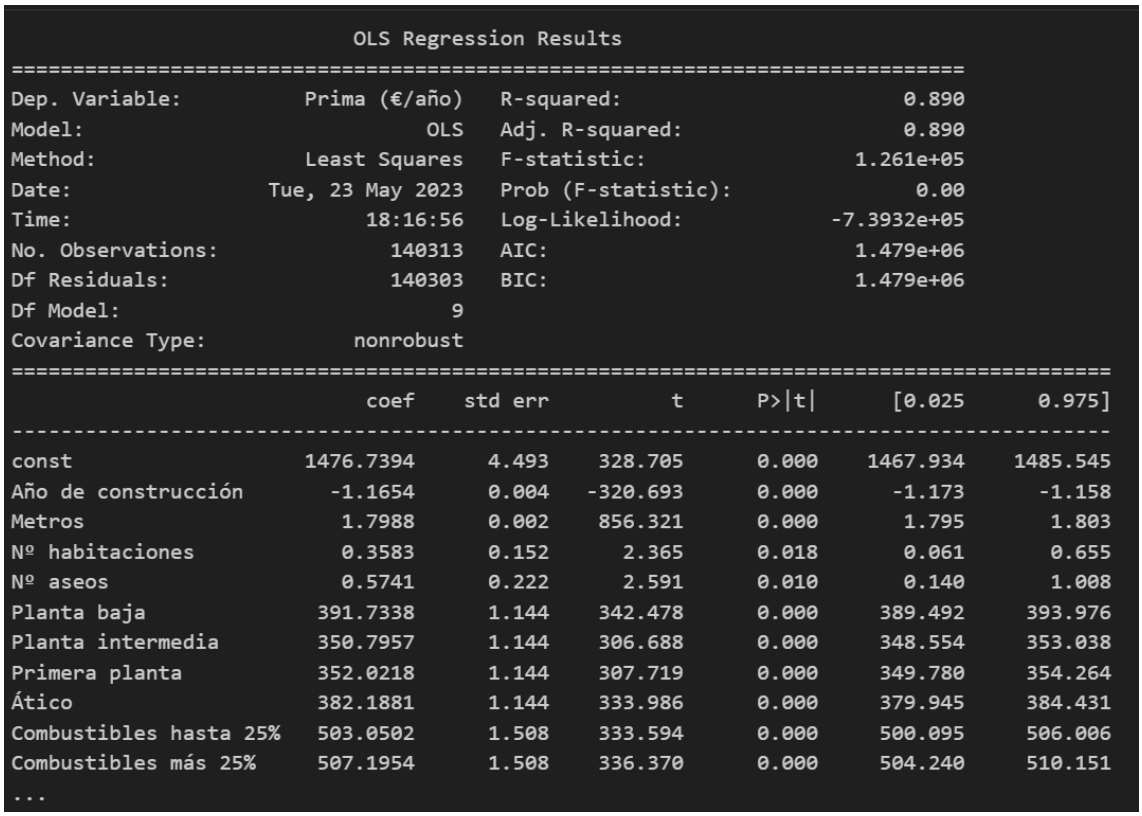
VALOR DE RANDOM_STATE	ERROR CUADRÁTICO	COEFICIENTE DE DETERMINACIÓN
42	2.185,86	89,00
12	2.193,12	89,06
22	2.209,45	88,95
63	2.210,40	88,96
387	2.204,28	88,99
1234	2.209,84	89,01

Tabla 1: Comparación del error cuadrático y el coeficiente de determinación obtenidos por el modelo según distintos valores de random_state, Fuente: Elaboración propia.

Como sucede con el tamaño, la distribución de los datos en el conjunto de datos de entrenamiento tampoco supone un gran cambio en la precisión del modelo.

5.3 Evaluación de la relación entre las variables independientes y la variable dependiente

Por último, resulta interesante analizar cómo las distintas variables independientes seleccionadas para construir el modelo afectan a este. De esta forma, se van a emplear métodos de análisis estadístico para ver si cada una de ellas tiene un impacto significativo en la prima predicha. De esta forma, se va a realizar mediante el ajuste por **mínimos cuadrados ordinarios** (OLS por sus siglas en inglés). Python, a través de su librería statsmodels ofrece una función **OLS()** que permite realizar este ajuste. En la **Figura 13** a continuación se presentan los resultados obtenidos.



OLS Regression Results

Dep. Variable:	Prima (€/año)	R-squared:	0.890
Model:	OLS	Adj. R-squared:	0.890
Method:	Least Squares	F-statistic:	1.261e+05
Date:	Tue, 23 May 2023	Prob (F-statistic):	0.00
Time:	18:16:56	Log-Likelihood:	-7.3932e+05
No. Observations:	140313	AIC:	1.479e+06
Df Residuals:	140303	BIC:	1.479e+06
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1476.7394	4.493	328.705	0.000	1467.934	1485.545
Año de construcción	-1.1654	0.004	-320.693	0.000	-1.173	-1.158
Metros	1.7988	0.002	856.321	0.000	1.795	1.803
Nº habitaciones	0.3583	0.152	2.365	0.018	0.061	0.655
Nº aseos	0.5741	0.222	2.591	0.010	0.140	1.008
Planta baja	391.7338	1.144	342.478	0.000	389.492	393.976
Planta intermedia	350.7957	1.144	306.688	0.000	348.554	353.038
Primera planta	352.0218	1.144	307.719	0.000	349.780	354.264
Ático	382.1881	1.144	333.986	0.000	379.945	384.431
Combustibles hasta 25%	503.0502	1.508	333.594	0.000	500.095	506.006
Combustibles más 25%	507.1954	1.508	336.370	0.000	504.240	510.151

...

Figura 13: Ajuste por Mínimos Cuadrados Ordinarios del modelo construido, Fuente: Elaboración propia.

En este caso resultan interesantes los **p-valores**[12] obtenidos por las distintas variables independientes. Se puede observar que, en la mayoría de los casos este tiene un valor de **0**, excepto para dos variables. Las variables independientes *Nº de habitaciones* y *Nº aseos* han obtenido un p-valor de **0,018** y **0,010** respectivamente. No obstante, todos los valores resultan ser inferiores a **0,05**, por lo que todas las variables parecen tener un impacto en la prima calculada por el modelo.

[12] El p-valor indica si un resultado es estadísticamente significativo. Cada p-valor contesta a una hipótesis. En estadística se determinan dos escenarios: la Hipótesis 0 (H0) representa la hipótesis nula y la Hipótesis 1 (H1) representa la hipótesis de investigación (aquello que se quiere evaluar). De esta forma, se determina que si p-valor < 0,05, entonces se acepta H1, de lo contario se acepta H0, es decir que la variable medida no tiene significancia estadística dentro del marco en el que está siendo evaluada.

6 Resultados y conclusiones

Este trabajo ha permitido explorar en profundidad la Regresión Lineal, así como su potencial uso en el sector asegurador. De esta forma, el algoritmo modelado ha obtenido una **precisión del 89%** al predecir las primas de un seguro de hogar. Esto aporta indicios acerca de la posible utilidad de este tipo de modelos en el sector.

No obstante, cabe destacar que, aunque la precisión resulta alta, puede no ser el mejor algoritmo para resolver este tipo de problemas. Efectivamente, se ha visto en el apartado 4.1 de este trabajo que las predicciones para primas inferiores a los 600 € al año se ajustan bien al modelo. Sin embargo, cuando la prima supera este importe, el ajuste de los valores predichos es cada vez menor. Asimismo, se puede observar en la **Figura 6**, una tendencia exponencial. Por tanto, esto podría ser un indicador que otro tipo de algoritmo sería más conveniente y se ajustaría mejor a los datos con los que se está trabajando.

En relación con estos, se debe de puntualizar que se ha construido el modelo entorno a un conjunto de datos reducido y que no representa todo el conjunto de variables que entran en juego a la hora de calcular un seguro de hogar. Efectivamente, en el Trabajo de Fin de Grado análogo a este, titulado **Análisis sobre el Cálculo de Primas de Seguros Generales,** se explica cómo funciona este cálculo. De esta forma, faltan elementos para tener en cuenta como los gastos en los que incurre una aseguradora, el margen de beneficio que desea obtener u otros aspectos que conciernen al tomador como, por ejemplo, su salud financiera. Se trata de datos que no se pueden encontrar en internet, puesto que se encuentran protegidos. No obstante, podrían tener cierto impacto en el modelo que no se está teniendo en cuenta.

De esta forma, el modelo construido se encuentra limitado a la disponibilidad de datos que se han podido encontrar. De cara a la propia aseguradora, al disponer de todos estos datos, la construcción de un modelo mucho más complejo resulta más fácil, por lo que la aplicación de estas técnicas de Inteligencia Artificial para el cálculo de las primas de seguro podría producir resultados más interesantes.

6.1 Revisión de objetivos

De cara a los objetivos fijados al inicio de este trabajo se puede destacar lo siguiente de cada uno de ellos:

- **Entender cómo se calculan las primas de seguros de hogar:** este trabajo, junto con el Trabajo de Fin de Grado análogo **"Análisis sobre el Cálculo de Primas de Seguros Generales"**, ha permitido indagar en el sector asegurador, analizando las distintas variables que pueden intervenir a la hora de fijar el precio de una prima de seguros.
- **Entender cómo funcionan los algoritmos de Regresión Lineal y cómo usarlos para el cálculo de primas de seguros de hogar:** se ha obtenido una mayor comprensión acerca de cómo funciona la Regresión Lineal dentro del marco del Machine Learning y cómo puede ser usada para lograr el objetivo principal de este trabajo: calcular una prima de seguro. Asimismo, se ha visto que, por las características de los datos, que mezclan variables cuantitativas y

categóricas, se podría probar otro tipo de algoritmo de Machine Learning que se ajustara mejor.

- **Buscar y tratar los datos necesarios para realizar el cálculo y alimentar el modelo:** en este punto se ha trabajado especialmente ya que la obtención de datos acerca de primas de seguro no resulta una tarea sencilla por la falta de disponibilidad. Efectivamente se trata de datos privados de cada una de las aseguradoras que no suelen estar expuestas en bases de datos de libre acceso, por lo que se ha tenido que recurrir a los calculadores online de primas de seguros. Asimismo, como se ha mencionado en el punto anterior, la presencia de datos categóricos ha obligado a un tratamiento de estos.
- **Analizar los resultados obtenidos y contrastarlos con resultados reales:** se han analizado los resultados obtenidos y contrastado con aquellos reales determinando una precisión del modelo razonable.

6.2 Futuros trabajos

En este trabajo se ha limitado el estudio a un único tipo de algoritmo de Machine Learning: Regresión Lineal. No obstante, y como se ha visto en los trabajos analizados en el apartado 4.2, puede resultar interesante **ampliar el estudio a otros tipos de algoritmos**. De esta forma, se pueden comparar los desempeños de cada uno de ellos y estimar si existe alguno que ofrezca mejores resultados.

Asimismo, como se ha comentado anteriormente, se ha realizado el estudio en base a una muestra de todas las variables que pueden entrar en juego a la hora de determinar la prima de un seguro de hogar. Por tanto, podría ser interesante, en un futuro, **ampliar los datos empleados** incluyendo otros como los listados a continuación:

- **Sistemas de seguridad**: alarma, puerta blindada, rejas, etc.
- **Estatus del tomador** respecto de la vivienda: propietario, arrendatario, etc.
- **Uso de la vivienda**: habitual, segunda vivienda, vivienda de vacaciones, etc.
- **Otros tipos de viviendas**: chalé adosado, chalé pareado, chalé individual, casa rural.
- **Localización** de la vivienda: núcleo urbano, zona despoblada, etc.
- **Valoración del contenido**: si se desea asegurar también los objetos de valor presentes en la vivienda, indicar cuánto valen.
- **Nacionalidad** del tomador de la póliza: nacional o extranjero.
- **Edad** del tomador.
- **Información acerca de la vivienda**: ha sido reformada, se encuentra actualmente asegurada, se encuentra hipotecada, etc.
- **Situación económica** del tomador: si se encuentra en la lista de morosidad[13], por ejemplo.

[13] Se trata de una variable muy importante para las aseguradoras a la hora de ofertar una póliza de seguro. Las compañías de seguros suelen elaborar lo que denominan como un "score". Se trata de un ranking de los clientes basado en el nivel de impago que presentan.

- **Forma de aseguramiento**: total (la prima corresponde a lo pactado en la póliza) o parcial (la prima asciende al valor de lo que se ha perdido).

7 Análisis de Impacto

En este capítulo se destacarán también aquellas decisiones tomadas a lo largo del trabajo que tienen como base la consideración del impacto.

Este último capítulo busca resaltar el impacto en diferentes contextos que ha tenido y puede llegar a tener el Trabajo de Fin de Grado realizado.

Desde el punto de vista **personal**, cabe destacar dos grandes aportaciones a la autora. Por una parte, el desarrollo de un modelo de **Regresión Lineal** ha aportado una mejor comprensión acerca del funcionamiento de este algoritmo usado en el campo de la **Inteligencia Artificial** y del **Machine Learning**. De esta forma, se ha aprendido cuáles son los distintos pasos necesarios para desarrollar un modelo basado en este algoritmo, desde la recogida de datos necesarios, hasta la evaluación final del modelo, pasando por la manipulación y transformación de los datos. Asimismo, al programar en **Python**, se han obtenido nuevos conocimientos acerca de este lenguaje que van a resultar muy útiles de cara al mundo laboral.

Por otra parte, se han obtenido conocimientos acerca del **sector asegurador** y, más concretamente de los elementos principales que entran en juego a la hora de calcular una prima de seguro de hogar. Efectivamente, los seguros son elementos esenciales que forman parte de la vida de cualquier persona. Y es que, ante los **riesgos** es necesario protegerse. Por tanto, analizar de forma más detallada aquellos elementos que pueden influir en la prima de una póliza de hogar aporta un valor añadido de cara al futuro para la autora del trabajo.

Cambiando el punto de vista cabe destacar que el sector asegurador es clave para la **economía**. Se trata de un sector que permite **fomentar la resiliencia** de las empresas, autónomos y familias que se pueden encontrar en situaciones de pobreza[14]. Además, genera empleo dando un **96% de contratos fijos**[26] promoviendo así la calidad en la empleabilidad[15]. Su función protectora afecta tanto a la industria como a las viviendas. De esta forma, protege un tejido industrial estimado en **1,73 billones de euros**[16], así como **20,7 millones de viviendas**[17] [26]. La aplicación de la Inteligencia Artificial al sector asegurador puede promover su crecimiento, manteniendo su impacto positivo en la economía. Efectivamente, al disponer de herramientas que permitan calcular de forma más ajustada y precisa los riesgos asociados a una póliza, las compañías podrán ajustar mejor sus precios y márgenes, reduciendo así los riesgos en los que ellas mismas incurren al ofrecer distintas coberturas.

Desde el punto de vista **social** el sector promueve la igualdad de género empleando un **52% de mujeres**[18][26]. Asimismo, también impulsa la **educación financiera** [19] ya que numerosas compañías aseguradoras participan en distintos proyectos enfocados en este ámbito. Por ejemplo, el proyecto ***El Riesgo y yo*** creado por Unespa ha conseguido reunir a unas 40 aseguradoras en su IV edición. Por último, se trata de un sector que permite la **reducción de**

[14] ODS nº1: Fin de la Pobreza.
[15] ODS nº8: Trabajo Decente y Crecimiento Económico.
[16] ODS nº9: Industria, Innovación e Infraestructura.
[17] ODS nº11: Ciudades y Comunidades Sostenibles.
[18] ODS nº5: Igualdad de Género.
[19] ODS nº4: Educación de Calidad.

desigualdades puesto que cualquier persona puede tener acceso a un seguro[20]. Respecto de este aspecto, la aplicación de técnicas de Machine Learning, tanto para el cálculo de primas como para otros aspectos puede tener **efectos perjudiciales**. Efectivamente, una dependencia excesiva de la Inteligencia Artificial aplicada puede ser que "muchos colectivos acaben expulsados del sector asegurador solo porque no encajan en los parámetros establecidos en el algoritmo, sin que se efectúe una valoración efectiva y real del riesgo" [27].

Por último, desde el punto de vista **medioambiental**, cabe destacar que las compañías del sector se encentran comprometidas con el **Desarrollo Sostenible**. De esta forma, aplican una serie de medidas como el ahorro de energía o el auge de la digitalización para **reducir su huella de carbono**[21]. Así, se puede ver que, como se ha estudiado en este trabajo, no sólo se puede emplear técnicas basadas en Inteligencia Artificial de cara a la actividad aseguradora, sino que también a un nivel interno dentro de la propia compañía, permitiendo una mejor monitorización de los consumos, por ejemplo.

[20] ODS nº10: Reducción de las Desigualdades.
[21] ODS nº12: Producción y Consumo Responsables y ODS nº13: Acción por el Clima.

8 Bibliografía

[1] MordorIntelligence, «Mercado de Seguros de Vida y No Vida en España | 2022 - 27 | Participación, tamaño y crecimiento de la industria - Mordor Intelligence». [En línea]. Disponible en: https://www.mordorintelligence.com/es/industry-reports/life-and-non-life-insurance-market-in-spain—growth-trends-and-forecast-2020—2025#:~:text=El%20sector%20asegurador%20en%20la

[2] Fundación Mapfre, «seguro de vida». [En línea]. Disponible en: https://www.fundacionmapfre.org/publicaciones/diccionario-mapfre-seguros/seguro-de-vida/

[3] Allianz, «¿Qué son los seguros generales? | Diccionario de Seguros Allianz». [En línea]. Disponible en: https://www.allianz.es/descubre-allianz/mediadores/diccionario-de-seguros/s/que-son-los-seguros-generales.html

[4] F. Mapfre, «ramo». [En línea]. Disponible en: https://www.fundacionmapfre.org/publicaciones/diccionario-mapfre-seguros/ramo/

[5] IBM, «¿Qué es el aprendizaje supervisado? | IBM». [En línea]. Disponible en: https://www.ibm.com/es-es/topics/supervised-learning

[6] Amazon Web Services, «¿Qué es la Regresión lineal? - Regresión lineal - AWS». [En línea]. Disponible en: https://aws.amazon.com/es/what-is/linear-regression/

[7] L. Gonzalez, «Regresión Lineal Simple - Teoría». abril de 2018. [En línea]. Disponible en: https://aprendeia.com/algoritmo-regresion-lineal-simple-machine-learning/

[8] C. Fernández-Simal Bernard, «Análisis sobre el Cálculo de Primas de Seguros Generales», 2023.

[9] Docs.Python.org, «Tutorial de Python — documentación de Python - 3.9.6». [En línea]. Disponible en: https://docs.python.org/es/3/tutorial/

[10] Á. Aguinaga, «Lenguajes de programación de bajo nivel VS alto nivel». abril de 2020. [En línea]. Disponible en: https://cipsa.net/lenguajes-de-programacion-de-bajo-nivel-vs-alto-nivel/#:~:text=Los%20como%20el%20c%C3%B3digo%20m%C3%A1quina

[11] IBM, «Programación orientada a objetos». [En línea]. Disponible en: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=language-object-oriented-programming

[12] InGenio Learning, «¿En qué se relaciona Python con la inteligencia Artificial?» abril de 2021. [En línea]. Disponible en: https://ingenio.edu.pe/blog/en-que-se-relaciona-python-con-la-inteligencia-artificial/#:~:text=%2C%20cursos%2C%20etc.-

[13] W. Schools, «Seaborn». [En línea]. Disponible en: https://www.w3schools.com/python/numpy/numpy_random_seaborn.asp

[14] Redacción, «Statsmodels : todo acerca de la biblioteca de Python». mayo de 2022. [En línea]. Disponible en: https://datascientest.com/es/statsmodels-todo-acerca

[15] A. de Madrid, «Censo de Edificios y Viviendas 2011». [En línea]. Disponible en: https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Edificacion-y-vivienda/Censo-de-edificios-y-viviendas/Censo-de-Edificios-y-Viviendas-2011/?vgnextfmt=default&vgnextoid=24fddc5bed1b8410VgnVCM1000000b205a0aRCRD&vgnextchannel=f93124e8951ef310VgnVCM1000000b205a0aRCRD

[16] Acierto.com, «¿Qué factores influyen en el precio del seguro de hogar?» [En línea]. Disponible en: https://www.acierto.com/seguros-hogar/precio-seguro-hogar/

[17] BBVA, «Cómo se calcula el precio de un seguro de hogar». mayo de 2023. [En línea]. Disponible en: https://www.bbva.es/finanzas-vistazo/ef/seguros/precio-seguro-hogar.html

[18] Caser, «Seguro de Hogar - Factores que influyen en el precio | Caser». [En línea]. Disponible en: https://www.caser.es/seguros-de-hogar/articulos/factores-que-determinan-el-precio

[19] Selectra, «¿Cuánto cuesta un seguro de hogar?» [En línea]. Disponible en: https://selectra.es/seguros/seguros-hogar/precios-seguros-hogar

[20] Catalano Occidente Seguros, «Cuánto cuesta un seguro de hogar y qué factores influyen». mayo de 2022. [En línea]. Disponible en: https://www.seguroscatalanaoccidente.com/blog/calcular-precio-seguro-hogar/

[21] M. Guillen y J. Pesantez-Narvaez, «Machine Learning y Modelización Predictiva para la Tarificación en el Seguro de Automóviles», *Anales del Instituto de Actuarios Españoles*, vol. 4ª época, pp. 123, 147, may 2018, doi: 10.26360/2018_6.

[22] Fundación Mapfre, «¿En qué consiste el fraude en seguros?» [En línea]. Disponible en: https://segurosypensionesparatodos.fundacionmapfre.org/seguros/definicion-seguro-asegurar/fraude-en-el-seguro-asegurar-riesgos/

[23] Allianz, «¿Qué es el Fraude en los seguros? | Diccionario de Seguros Allianz». [En línea]. Disponible en: https://www.allianz.es/descubre-allianz/mediadores/diccionario-de-seguros/f/fraude.html

[24] E. B. Valero, A. S. Díaz, y J. S. Gisbert, «Algoritmos de Machine Learning para la Detección del Fraude en el Seguro de Automóviles», *Anales del Instituto de Actuarios Españoles*, vol. 4ª época, pp. 23, 46, may 2020, doi: 10.26360/2020_2.

[25] L. de los A. M. Asencio Diaz, R. H. Chiang Cornejo, F. L. Crisóstomo Fernández, G. V. Hernández Quiroz, y A. S. Lajo Aurazo, «Técnicas de Machine Learning para la clasificación automática de clientes en una empresa de seguros», 2021. [En línea]. Disponible en: https://repositorio.esan.edu.pe/handle/20.500.12640/2933

[26] M. Larrauri, «Los Objetivos de Desarrollo Sostenible en el seguro | Blog Estamos Seguros». mayo de 2023. [En línea]. Disponible en: https://www.estamos-seguros.es/los-objetivos-de-desarrollo-sostenible-en-el-seguro/

[27] V. M. Zamarreño, «Inteligencia Artificial para hacer al seguro un poco más humano». elEconomista, mayo de 2023. [En línea]. Disponible en: https://www.eleconomista.es/actualidad/noticias/12180656/03/23/Inteligencia-Artificial-para-hacer-al-seguro-un-poco-mas-humano.html

9 Anexos

A continuación, se presenta el código completo en lenguaje Python en base al cual se ha elaborado el trabajo

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf


from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.preprocessing import OneHotEncoder
from scipy.stats import pearsonr
from scipy import stats


data = pd.read_csv('insurance 2.csv', sep= ';')
data.dtypes
```

Ilustración 1: Carga de librerías, así como del dataset, Fuente: Elaboración propia.

```
data["Prima (€/año)"] = data["Prima (€/año)"].str.replace(',', '.')
data = data.astype({'Prima (€/año)': float})

data["Nº habitaciones"] = data["Nº habitaciones"].replace('6+', '6')
data = data.astype({'Nº habitaciones': int})

data["Nº aseos"] = data["Nº aseos"].replace('3+', '4')
data = data.astype({'Nº aseos': int})

data.dtypes
```

Ilustración 2: Transformación de los datos a un formato numérico, Fuente: Elaboración propia.

```
plt.scatter(data['Metros'], data['Prima (€/año)'], color='red')
plt.scatter(data['Año de construcción'], data['Prima (€/año)'], color='blue')
plt.scatter(data['Código Postal'], data['Prima (€/año)'], color='green')
plt.scatter(data['Nº habitaciones'], data['Prima (€/año)'], color='pink')
plt.scatter(data['Nº aseos'], data['Prima (€/año)'], color='orange')
```

Ilustración 3: Código para graficar las variables independientes y comprobar la linealidad, Fuente: Elaboración propia.

```
X=data.iloc[:,2:8]
y=data.iloc[:,8]
```

Ilustración 4: Determinación del conjunto de variables independientes (X) y de la variable dependiente (y), Fuente: Elaboración propia.

```
X= pd.get_dummies(data=X, drop_first=True)
X.head()
```

Ilustración 5: Transformación de las variables categóricas mediante el uso de variables dummies, Fuente: Elaboración propia.

```
ohe = OneHotEncoder(sparse=False, handle_unknown='ignore')
X_cat = pd.DataFrame(ohe.fit_transform(data[['Tipo de vivienda', 'Materiales']]))

ohe.categories_

X_final = X.join(X_cat)
X_final.columns = ['Año de construcción', 'Tipo de vivienda', 'Materiales', 'Metros', 'Nº habitaciones', 'Nº aseos',
                   'Planta baja', 'Planta intermedia', 'Primera planta', 'Ático', 'Combustibles hasta 25%', 'Combustibles más 25%', 'Incombustibles']
X_final.drop('Tipo de vivienda', axis = 1, inplace=True)
X_final.drop('Materiales', axis = 1, inplace=True)

X = X_final
```

Ilustración 6: Transformación de las variables categóricas por el método One Hot Encoder, Fuente: Elaboración propia.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("X_train: ", X_train.shape)
print("X_test: ", X_test.shape)
print("y_train: ", y_train.shape)
print("y_test: ", y_test.shape)
```

Ilustración 7: Separación de los datos entre el conjunto de entrenamiento y de prueba, Fuente: Elaboración propia.

```
model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
print('Error cuadrático:', mse)

accuracy=r2_score(y_test,y_pred)*100
print("La precisión del modelo es %.2f" %accuracy)

plt.scatter(y_test, y_pred)
plt.xlabel('Prima Real')
plt.ylabel('Prima Estimada')
plt.title('Prima Real vs Estimada')
plt.show()
```

Ilustración 8: Entrenamiento del modelo y evaluación de las predicciones, Fuente: Elaboración propia.

```
sns.regplot(x=y_test, y=y_pred, ci=None, color='blue')
pred_df = pd.DataFrame({'Actual Value': y_test, 'Predicted Value': y_pred, 'Difference': y_test-y_pred})
```

Ilustración 9: Código para graficar las predicciones en función de los valores reales junto con la recta de regresión, Fuente: Elaboración propia.

```
X_train = sm.add_constant(X_train, prepend=True)
modelo = sm.OLS(endog=y_train, exog=X_train,)
modelo = modelo.fit()
print(modelo.summary())
```

Ilustración 10: Evaluación del modelo y del nivel de significación de las variables independientes, Fuente: Elaboración propia.

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Tue May 30 17:32:12 CEST 2023
	Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
	Numero de Serie	561
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)