Islas Canarias
Del 15 al 19 de noviembre de 2021

UNA HERRAMIENTA DE MACHINE LEARNING PARA LA ACTUALIZACIÓN Y EL DESARROLLO DEL DIRECTORIO DE EMPRESAS Y ESTABLECIMIENTOS CON ACTIVIDAD ECONÓMICA EN ANDALUCÍA

Nuria Gómez-Vargas
Instituto de Matemáticas de la Universidad de Sevilla
ngvargas@us.es

Jasone Ramírez-Ayerbe
Instituto de Matemáticas de la Universidad de Sevilla
mrayerbe@us.es

Emilio Carrizosa Priego
Instituto de Matemáticas de la Universidad de Sevilla
ecarrizosa@us.es

Rafael Blanquero Bravo
Instituto de Matemáticas de la Universidad de Sevilla
rblanquero@us.es

Marina Enguídanos Weyler
Instituto de Estadística y Cartografía de Andalucía (IECA)
marina.enguidanos@juntadeandalucia.es

Elisa Isabel Caballero Ruiz
Instituto de Estadística y Cartografía de Andalucía (IECA)
elisai.caballero@juntadeandalucia.es

Ana Gema Galera Pozo
Instituto de Estadística y Cartografía de Andalucía (IECA)
gema.galera@juntadeandalucia.es

Introducción

Nuestro trabajo se encuentra en el Marco de un conjunto de actuaciones de transferencia en los campus de excelencia internacional en torno a la inteligencia artificial con aplicación a los sectores RIS3[1]. El citado Marco regula que las universidades presentarán un conjunto de proyectos entorno a la Inteligencia Artificial con aplicación a los sectores RIS3. Se trata de un proyecto en colaboración entre el Instituto de Estadística y Cartografía de Andalucía y el Instituto de Matemáticas de la Universidad de Sevilla, cuyo objetivo es la caracterización del tejido empresarial andaluz y en el que podemos distinguir dos partes.

Por un lado, tenemos la tarea de identificación de empresas y establecimientos con actividad en Andalucía. Actualmente, la actualización del Directorio se realiza integrando la información procedente de registros administrativos (Impuesto de Actividades Económicas y Seguridad Social) y fuentes complementarias (Registro Mercantil, Guía de la Banca). Como las fuentes utilizadas para esta actualización tienen características y fechas de disponibilidad diferentes, se requiere un tratamiento particularizado, que hace más costosa la elaboración del Directorio. Habría que reducir esta complejidad identificando fuentes de información complementarias e implementado procedimientos de integración de las mismas. Es por ello por lo que se propone una metodología alternativa. Gracias a la información que se encuentra en la web, se propone la identificación de fuentes de información alternativas y la captura automática de su contenido, haciendo uso de *web scraping* mediante librerías disponibles para Python.

Y, por otro lado, hemos abordado un análisis del carácter innovador de las empresas. Actualmente, el Instituto Nacional de Estadística ofrece información directa sobre el proceso de innovación en las empresas de más de diez empleados en su "Encuesta sobre innovación en las empresas", elaborando indicadores que permiten conocer los distintos aspectos de este proceso. No obstante, quedan excluidas de esta encuesta las empresas de menos de diez empleados, las cuales representan en Andalucía el 95% del total. Es por ello que proponemos una metodología alternativa para caracterizar la innovación de cualquier empresa: el desarrollo de un modelo de aprendizaje automático implementado en Python que permita clasificar a las empresas según su innovación a partir de variables extraídas de sus páginas webs. Con esto no solo conseguiremos discriminar a las empresas según su carácter innovador sino también acabar dando una definición para la innovación en Andalucía a partir de las variables importantes.

Identificación de Empresas y Establecimientos en Andalucía

Objetivos

El objetivo principal de esta parte es actualizar y caracterizar el Directorio de Empresas y Establecimientos en Andalucía. Como objetivos específicos, primero es necesario identificar fuentes alternativas donde se almacene información sobre empresas y establecimientos en Andalucía, y que está pueda ser capturada automáticamente. Una vez identificadas las fuentes, se pretende llevar a cabo la captura automática de su información. Esto es posible gracias a las técnicas de *web scraping* basándonos en las

[1] Actuaciones en Áreas de la estrategia de investigación e innovación para la especialización inteligente de Andalucía (RIS3), las cuales están cofinanciadas por el Programa Operativo FEDER en Andalucía para el periodo 2014-2020.

librerías que están disponibles en Python, específicamente *Selenium*. Una vez que se ha obtenido la información, se debe realizar un análisis de calidad de los datos. Más concretamente, se debe comprobar la calidad de la dirección postal y su enlace con la información existente en el Directorio. Si la calidad es suficientemente buena, con la ayuda de Python podremos actualizar el directorio, detectando altas y bajas o actualizando variables, como el número de teléfono, la página web, actividad económica, coordenadas geográficas de ubicación u otros datos de la empresa/establecimiento.

Metodología

El punto de partida de este estudio son los diferentes recursos de información que se encuentran en Internet y que cumplen con los requisitos definidos por el IECA. Se deben poder definir búsquedas limitadas que filtren automáticamente solo los datos solicitados. Todo este procedimiento se realiza mediante técnicas de *web scraping*; específicamente se ha escogido trabajar en Python, con la librería *Selenium*, que nos da suficiente libertad y flexibilidad para trabajar tanto con páginas estáticas como dinámicas. *Selenium* controla remotamente las instancias de los navegadores, con lo que se pueden emular las posibles interacciones de los usuarios. Permite así completar campos de texto, seleccionar opciones de desplegables, hacer click, etc. de una manera intuitiva y fácil de codificar. Aún así, el código de *web scraping* depende siempre de las etiquetas específicas de la página web, con lo que debe de escribirse un código por cada fuente de información. Por esta razón se escogen páginas web en las que se encuentre la máxima información posible y cuya estructura y configuración sea además estable.

Es preciso realizar una distinción entre el procedimiento para la captura de información de establecimientos y de empresas.

En el caso de las empresas, se decide "escrapear" la página web eInforma. eInforma es una web estática donde se encuentra información de empresas según su CNAE, lo que nos permite realizar un filtrado según la actividad económica que se desee, y además se puede filtrar según la provincia en la que se encuentra, pudiéndonos así centrar en Andalucía. Para comprobar la calidad de los datos y realizar los enlaces con la información del directorio del IECA, se realiza primero un enlace directo según el NIF de la empresa, dato que es scrapeado sin problema. Los enlaces de esta forma son directos y fiables. En el caso de las no coincidencias, se observa que estas disminuyen cuando se verifican los establecimientos inactivos; además se puede realizar una evaluación más amplia debido al *web scraping* masivo, inspeccionando empresas que han sido recuperadas, pero con una CNAE asociada a otro código dentro del mismo sector. En el caso de las coincidencias, se analiza la calidad de la fuente comparando las direcciones asociadas a la misma empresa. Se asigna un valor entre 0 y 1 a cada coincidencia, siendo 0 la distancia mínima y el mejor de los casos. Se asigna pues un valor de 1 si ni siquiera coinciden los códigos postales; si, por el contrario, coinciden, entonces se calcula la distancia entre la cadena de caracteres de las dos direcciones. Existen diferentes distancias para comparar cadenas literales, como Levenshtein o Jaro-Winkler, siendo esta última más acertada. Para la comparación de cadenas primero hacemos un preprocesamiento, eliminando de cada registro las llamadas "stopwords", que no son más que las palabras más comunes de un idioma como preposiciones o artículos.

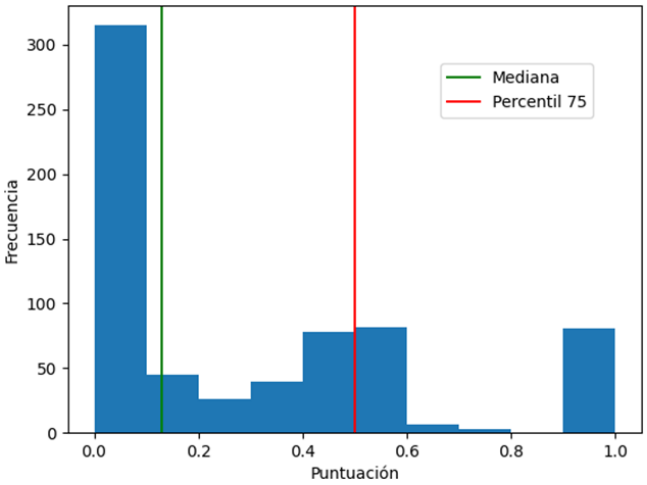
En el caso de establecimientos nos centramos en el sector bancario. Nos encontramos dos tipos de fuentes distintas. Por una parte, se encuentran directorios de establecimientos

estáticos como Credimarket, Iahorro o Sucursales.org. En estos se encuentra información sobre la dirección de los locales de las entidades bancarias en Andalucía, si bien hay que realizar un análisis de los datos exhaustivo. Para este caso, se crea nuevamente un código en Python para cada una de las páginas, se "escrapea" la información y se realizan los enlaces primero de la entidad bancaria, utilizando el NIF del banco si es uno de los datos extraídos o bien haciendo uso del literal del nombre y realizando su comparación con la del directorio. Una vez encontrada esta coincidencia, se comparan los literales de las direcciones, asignándose al más cercano. Por otra parte, se realiza el "escrapeo" de la página web OpenStreetMap. En el caso de OpenStreetMap, se obtiene la posición geográfica exacta de los establecimientos. Gracias a esto, utilizando la Fórmula de Haversine, podemos calcular la distancia entre dos puntos y vincular las direcciones del *web scraping* a la más cercana del directorio. Asignamos una puntuación de coincidencia de 1, 0.5 o 0, según una serie de condiciones.

Una vez que se ha realizado el enlace, se pasa a la actualización de variables (nº de teléfono, fax, página web, …). En el caso de las empresas, eInforma incluye la información de altas y bajas, con lo que podemos actualizar está información una vez se ha enlazado la información.

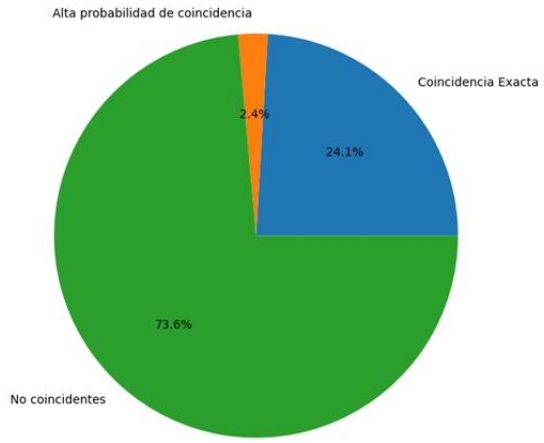
Resultados

Podemos ver los resultados del scrapeo de empresas y establecimientos. En cuanto a las empresas, nos centramos en aquellas con CNAE 1043, el sector de fabricación de aceite de oliva; se obtienen 1044 empresas, más que las que había en el directorio, que eran 996. Se realiza el enlace y se comprueban las puntuaciones. Podemos ver en un histograma de las distancias que: la gran mayoría se encuentran a distancia 0, que no es más que la coincidencia absoluta, y el percentil 75 es 0.5. Las diferencias son principalmente de pequeñas variaciones en la normalización de la dirección. Las no coincidencias pueden darse al haber un cambio de dirección. Además, en el caso de empresas se puede realizar un estudio de altas y bajas, comprobándose que la mayoría siguen de alta, aunque hay un pequeño porcentaje, 1.4% que pasan de alta a baja, una información relevante.



En cuanto a los establecimientos, se presenta aquí la información obtenida en OpenStreetMap (OSM), se puede observar que no cubre toda la información presente en el directorio, pero la que coincide lo hace de forma exacta. Por ejemplo, en el caso del sector bancario:

- Se localizan 1838 establecimientos bancarios en Andalucía bajo la etiqueta “bank”.
- Se localizan 347 establecimientos bancarios en Andalucía bajo la etiqueta “atm”.
- Se localizan 14 establecimientos bancarios en Andalucía bajo la etiqueta “bureau_de_change”.
- En total se localizan 2199 establecimientos bancarios, mientras que en el directorio del IECA hay 5752 establecimientos, estando activos- a fecha de 2019- 4379 de ellos.
- Una de las ventajas de OSM es que no hace falta normalizar las direcciones porque todas las instancias devuelven las coordenadas de latitud y longitud, por lo que se puede usar el servicio web del Callejero Digital de Andalucía Unificado (CDAU) “geocoderInverso” para obtener la dirección.



Esto nos muestra como en el caso de establecimientos no es posible hacer uso de una sola fuente, sino que es necesaria la complementación de varias.

Análisis del carácter innovador de las empresas

Objetivos

Como objetivo principal de esta parte del proyecto, tenemos el análisis del carácter innovador de las empresas. Para ello, nos marcamos como objetivos específicos el desarrollo de un modelo de *machine learning* o aprendizaje automático que nos permita clasificar a empresas a partir de una muestra etiquetada según si son innovadoras o no. Además, del análisis de este modelo queremos definir el concepto de innovación en Andalucía, por lo que extraemos las variables importantes y hacemos estudios marginales. Finalmente, deseamos comprobar si este concepto y el mismo modelo es extrapolable para caracterizar a las empresas de menos de diez empleados según su innovación.

Metodología

Como introdujimos al principio, queremos clasificar a las empresas según variables extraídas de sus páginas webs. Para extraer estas variables descriptivas hacemos uso de técnicas de *web scraping*, concretamente usando de manera conjunta las librerías *requests* y *BeautifulSoup* de Python. Elegimos Python fundamentalmente por ser un lenguaje intuitivo y por las librerías tan completas que se han desarrollado para llevar a cabo esta técnica.

Para ello nos basamos en la metodología (Daas et al. 2020) para *escrapear* el texto del dominio web de una empresa del Instituto de Estadística Holandés (CBS). Consiste en la descarga del contenido de la página haciendo la petición con las funciones de la librería *requests*. Este contenido se le pasa posteriormente a *BeautifulSoup* para extraer los datos del código html y generar un árbol de análisis con el que podremos seleccionar las variables que nos interesan para nuestro modelo. Aquí es donde ampliamos la metodología anterior, ya que nuestras variables no solo van a ser el texto extraído sino también variables sobre cómo ha sido construido el sitio web (longitudes de la cabecera y el cuerpo html, meta etiquetas, número de imágenes que contiene o idiomas que podemos encontrar).

Es necesario hacer un preprocesamiento con minería de texto de la "sopa" extraída con BeautifulSoup. Para ello, hay también implementadas en Python librerías de procesamiento del lenguaje natural, como *nltk*. Con ella hacemos una limpieza de los caracteres (eliminando símbolos y números), eliminación de *stopwords* (en español encontramos, por ejemplo, artículos y pronombres) y también hacemos lo que en inglés se conoce como *stemming*, que consiste en quedarnos sólo con la raíz de las palabras, porque lo que nos interesa es la semántica.

Debido al gran volumen de variables, tenemos que implementar técnicas de selección y agrupación. Para ello, decidimos formar clústers de palabras según el peso que estas tendrían en un modelo de clasificación logística penalizada y con la innovación como objetivo. Seleccionamos como técnica de agrupación el *clustering* por densidades con el algoritmo DBSCAN, porque no nos obliga a predefinir un número determinado de clústers y además no tiene por qué agrupar todas las variables, dejando fuera de los grupos a palabras *noise*. Esta técnica la encontramos junto con otros algoritmos para el análisis y predicción a partir de datos en la librería *sklearn*. Scikit-learn es la librería más útil para *machine learning* en Python, y proporciona algoritmos de aprendizaje supervisado y no supervisado.

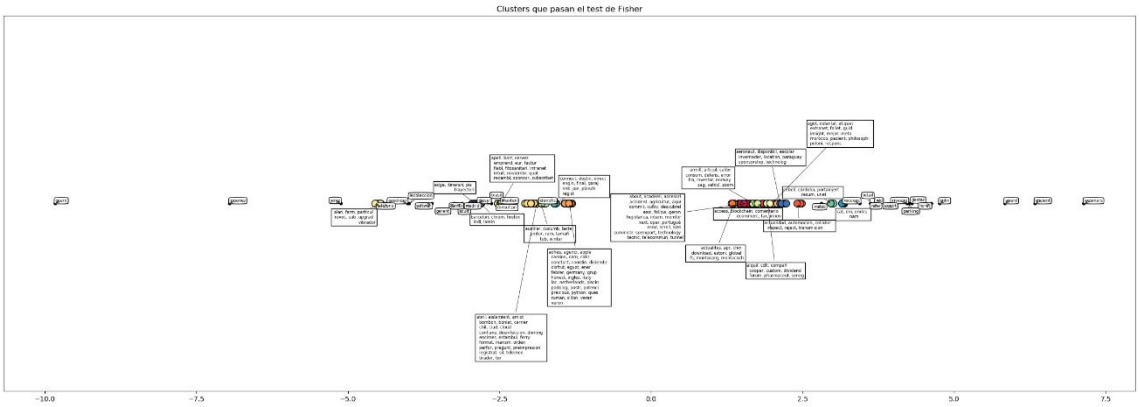
Además, también hacemos una selección de aquellas palabras y etiquetas cuyas apariciones en la web muestran dependencia con el carácter innovador de estas. Esta información nos la proporciona el Test de Fisher, que está ya implementado en la librería *scipy*- muy utilizada en las áreas de matemáticas e ingeniería.

Una vez que ya tenemos definida nuestra matriz de variables construimos el modelo de clasificación. La librería *sklearn* no sólo nos permite implementar el modelo Random Forests sino que también tiene implementado el *tuning* de sus parámetros (número de árboles, profundidad, número de variables seleccionadas en cada nodo, etc.) mediante técnicas tales como la validación cruzada. El resto de los parámetros que hemos ido definiendo en nuestra metodología completa y que necesitan ajustarse se han tuneado con optimización Bayesiana haciendo uso de la librería GPyOpt. Esta opción es de las más utilizadas para optimizar modelos de *machine learning*, ya que funciona incorporando información aprendida en evaluaciones de funciones anteriores para elegir un conjunto óptimo de coordenadas para la siguiente evaluación, permitiendo así no tener que evaluar todas las posibles combinaciones de los *tuning parameters*.

Resultados

Las salidas de nuestro modelo nos permiten ir haciendo el análisis que queríamos sobre la innovación en Andalucía. Hemos conseguido entrenar este modelo con tan sólo 87

variables explicativas, habiendo partido de una base de datos con 1310 empresas y 4336 variables.



El primer gráfico que nos encontramos son los clústers de palabras significativos con los que nos quedamos, que son 30 palabras noise y 18 grupos. Destacamos de este gráfico los *outliers*, las palabras que influyen de manera distinta a las demás en el carácter innovador:

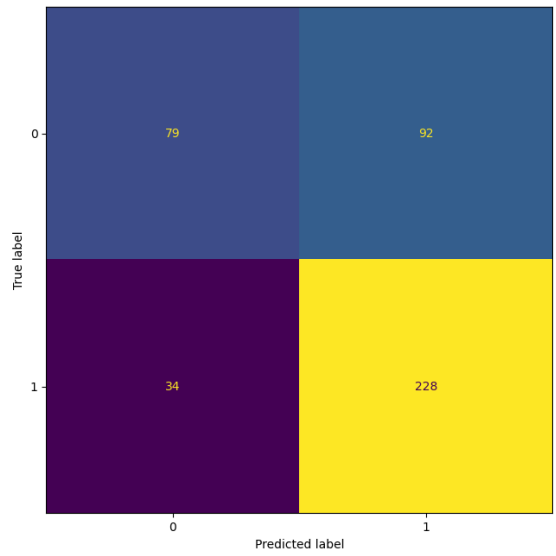
- Influyen con un peso positivo en la clasificación, es decir, clasifican hacia un carácter innovador, palabras como “webinars”, “eficiente”, “online”, “award”, “investigación”, “destacar”.

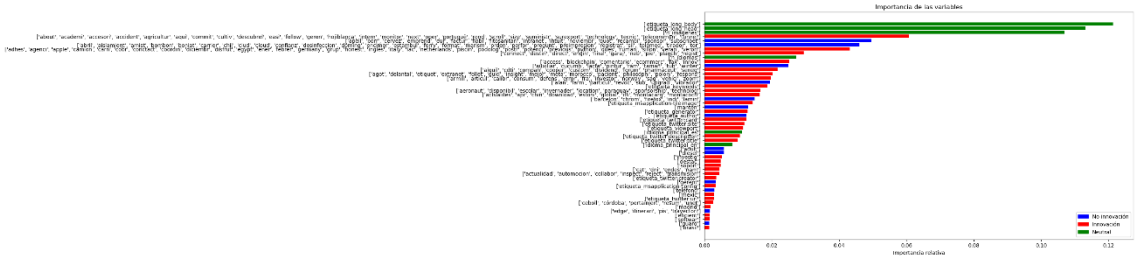
- Influyen con un peso negativo en la clasificación, es decir, clasifican hacia un carácter no innovador, palabras como “jimenez”, “jerez”, “recolección”, “mantenimiento”, “consultar”, “teléfono”, “gasolinera”.

En cuanto a los clústers formados:

- Comprobamos que no hay clusters con pesos en torno al cero, resultado de usar una clasificación logística penalizada y que efectivamente esas palabras no fueran significativas.
- En el lado derecho encontramos clústers de palabras que aumentan el carácter innovador como {acceso, blockchain, comentario, ecommerce, fax, innovación}.
- En el lado izquierdo encontramos clústers de palabras que disminuyen el carácter innovador como {apellido, cerveza, euro, factura, intranet, subscripción, …}.

El modelo de clasificación final construido con mil árboles nos da una precisión obtenida por validación cruzada del 70%. De la matriz de confusión, obtenida sobre una evaluación de la muestra test, podemos observar que nuestro modelo es capaz de identificar la innovación correctamente, pero que la no innovación no debe ser tan homogénea y resulta más difícil de capturar.





La técnica de *Random Forests* no sólo ha sido elegida por su poder predictivo, sino también porque muestra un análisis de las variables más importantes en la clasificación. Podemos ver en verde las variables que son etiquetas de conteo o idiomas, en rojo las variables que son significativas para la innovación, y en azul las que son significativas para la no innovación.

- Las tres variables más importantes, en orden, son: longitud del cuerpo html (long_body), longitud de la cabecera html (long_head), número de imágenes en la web. A pesar del extenso estudio de las palabras, lo más importante a la hora de discernir es cómo está construida la propia web. Puede deberse a que realmente esto sea decisivo o a la ambigüedad en las definiciones de innovación o no innovación.

- Vemos una mayor cantidad de barras rojas que azules, de ahí que la innovación se detecte mejor.

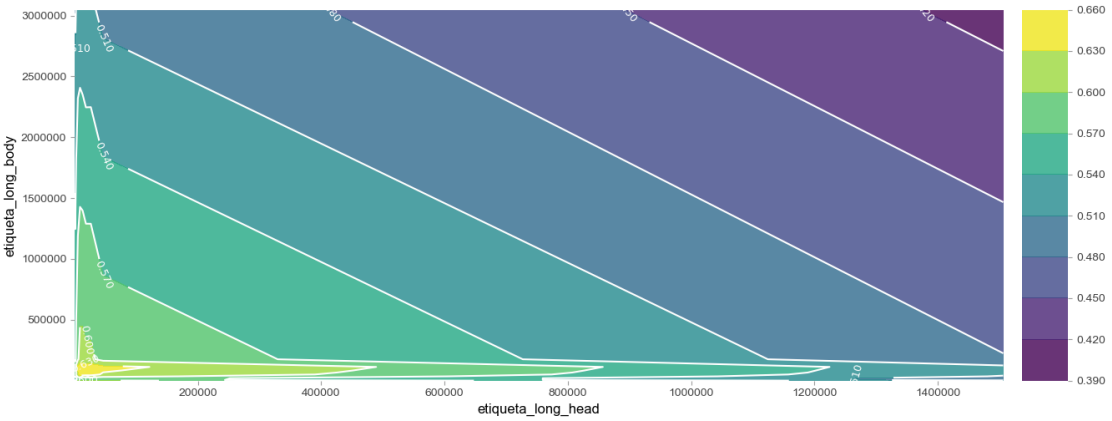
- Las siguientes variables importantes son clústers, luego la agrupación por pesos ha tenido sentido.

    - El primero que encontramos que es significativo para la innovación es {'about', 'academi', 'accesori', 'accident', 'agricultur', 'aquí', 'commit', 'cultiv', 'descubrel', 'easi', 'follow', 'geron', 'hojiblanca', 'intern', 'monitor', 'next', 'oper', 'portuguê', 'rond', 'scroll', 'size', 'suministr', 'surexport', 'technology', 'tecnic', 'telecommun', 'tunnel'}.
    - El primero que encontramos que es significativo para la no innovación es {'apell', 'borr', 'cervez', 'emprend', 'eur', 'factur', 'fiabl', 'fitosanitari', 'intranet', 'intuit', 'noviembr', 'quot', 'recambi', 'sponsor', 'subscribet'}.

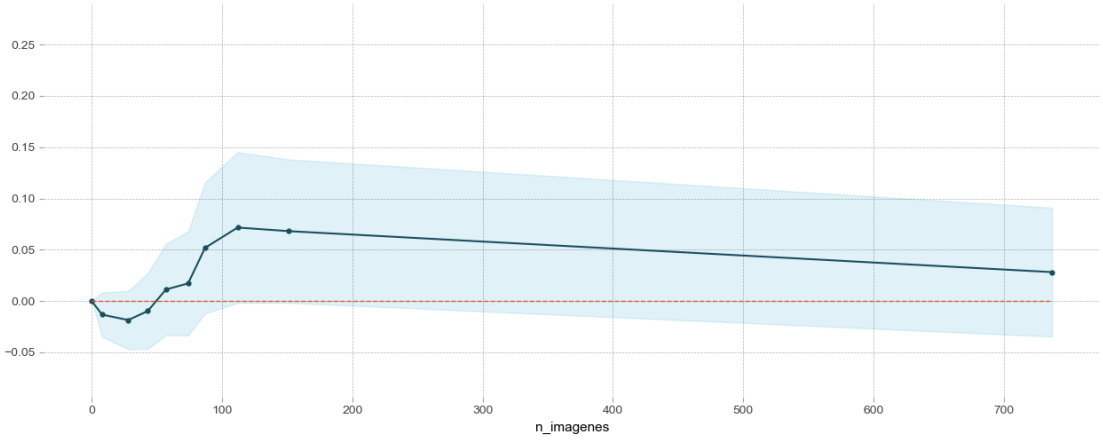
- A pesar de la gran variedad de idiomas que encontramos en las webs, lo importante es la cantidad de idiomas que presenta, más que cuál específicamente. Lo único que predomina es si la web está en inglés o en español.

Finalmente, obtenemos gráficos de la influencia marginal de las variables más importantes en la decisión sobre la clasificación:

- Tiene sentido estudiar la interacción de las variables 'long_head' y 'long_body' antes que sus efectos marginales, pues en una web las encontramos obligatoriamente siempre a las dos. Podemos comprobar la relación de manera inversamente proporcional de la longitud con el carácter innovador, a medida que aumentan estas longitudes también lo hace el carácter no innovador.



- En cuanto al gráfico del efecto marginal de la variable 'n_imagenes' sobre la clasificación, podemos observar que, si la página web no presenta apenas imágenes, o tiene demasiadas, se tiende hacia la no innovación.



Conclusiones

Respecto a la actualización del directorio, se concluye que la presencia de información de empresas es elevada, y además de buena calidad. Esto nos permite capturar la información automáticamente mediante *web scraping* y realizar la actualización automática de distintas variables, así como detectar nuevas altas y bajas de empresas. En cuanto a los establecimientos, la información que hay es menor, y sobre todo de peor

calidad. Se necesita un estudio complementario de varias fuentes y un análisis de calidad más exhaustivo.

Y respecto al análisis de la innovación, concluimos que las empresas que no son innovadoras en Andalucía tienen actividad en sectores muy variados, mientras que las empresas innovadoras son más homogéneas y de ahí que las detectemos mejor. Además, frente a estudios anteriores sobre la caracterización de empresas a partir de sus páginas webs- que se basan en una minería de texto de las mismas-, incorporamos variables relativas a la propia construcción y estructura de la web, lo que resulta tener una importancia destacable en cuanto a la clasificación. Por último, con la definición de innovación que obtenemos gracias al modelo de aprendizaje supervisado, podríamos hacer un estudio en el tiempo para analizar si el concepto cambia a partir de las variables importantes.