Islas Canarias
Del 15 al 19 de noviembre de 2021

imus instituto de matemáticas universidad de sevilla

Junta de Andalucía | Consejería de Transformación Económica, Industria, Conocimiento y Universidades | Instituto de Estadística y Cartografía de Andalucía

XXI Jornadas de Estadística de las Comunidades Autónomas

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, Marina Enguídanos Weyler

Nuria Gómez Vargas, Marina Enguídanos Weyler

Índice

Introducción

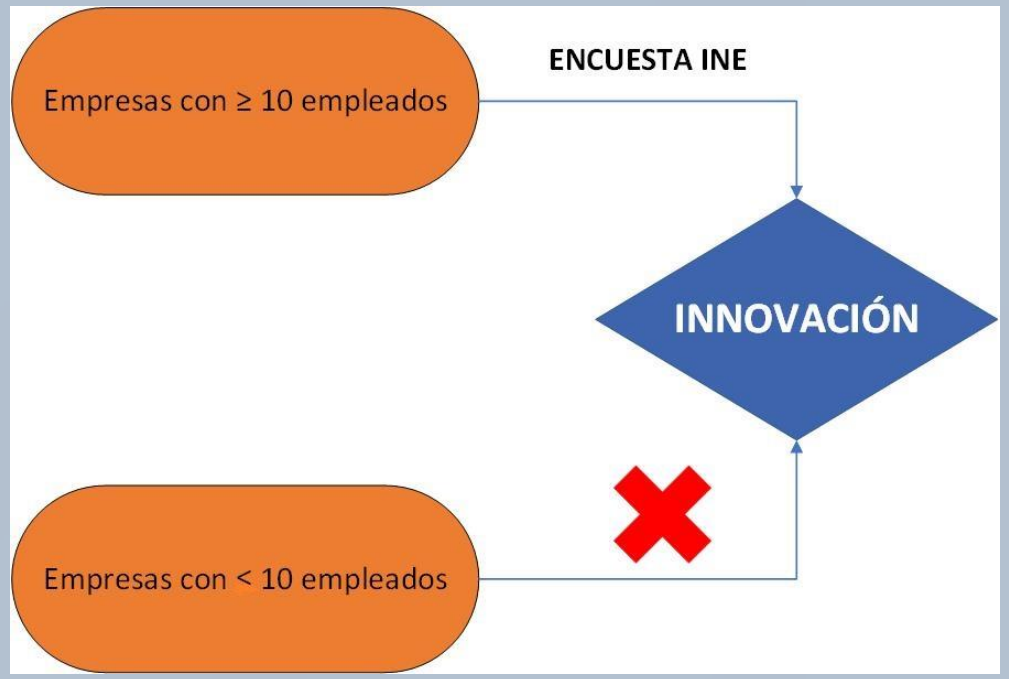
- **Identificación de Empresas y Establecimientos en Andalucía**

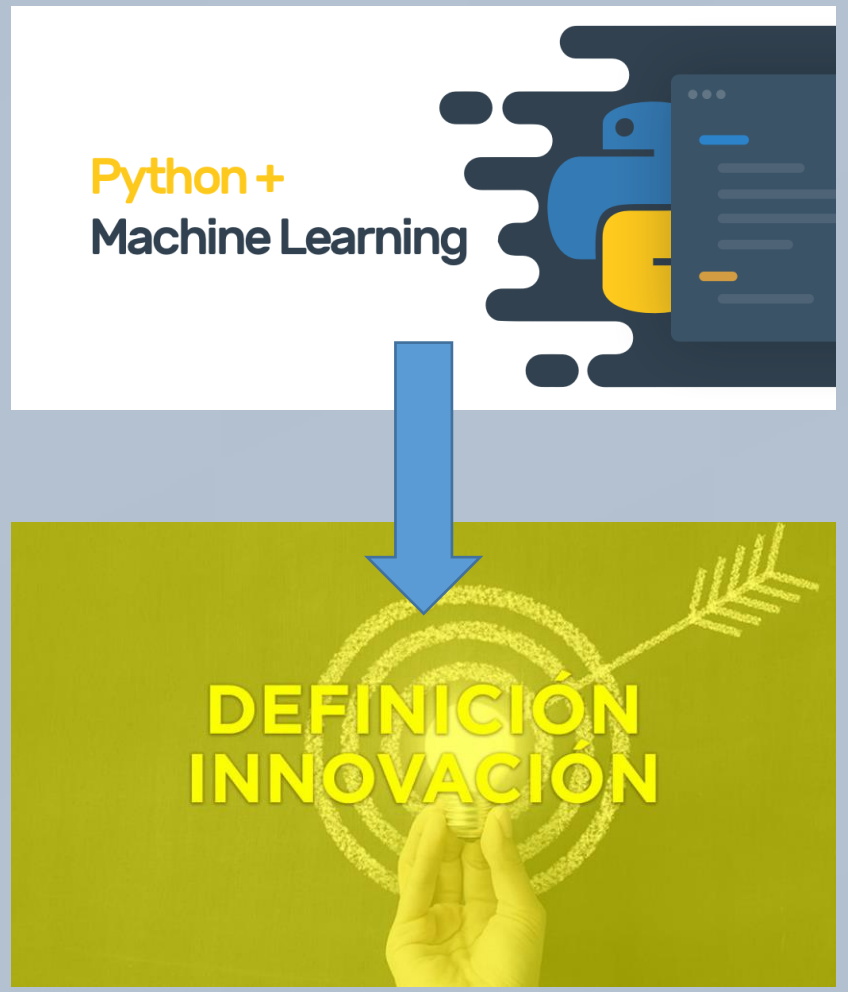
Tradicional

Nueva dinámica

Introducción

- **Análisis del carácter innovador de las empresas**





Identificación de Empresas y Establecimientos en Andalucía

Objetivos

Actualizar y contrastar el Directorio de Empresas y Establecimientos en Andalucía

1. Identificación de fuentes de información alternativas
2. Captura automática de la información: *web scraping*
3. Análisis de calidad
4. Actualización automática del Directorio con Python
   a) Detección altas y bajas
   b) Actualización dinámica de las variables

Identificación de Empresas y Establecimientos en Andalucía
Metodología

Web scraping: Python, Selenium

Código por página

```
driver=webdriver.Chrome()
driver.get(url)
links=[]
empresas=driver.find_elements_by_css_selector(".colorBlue")
provincias=driver.find_elements_by_css_selector(".first~ td+ td span")

while True:
    try:
        driver.find_element_by_partial_link_text("siguiente").click()
        empresas=driver.find_elements_by_css_selector(".colorBlue")
        provincias=driver.find_elements_by_css_selector(".first~ td+ td span")
        for i in range(len(provincias)):
            if provincias[i].text in andalucia:
                links.append(empresas[i].find_element_by_tag_name("a").get_attribute('href'))
    except:
        break
```

Empresas:

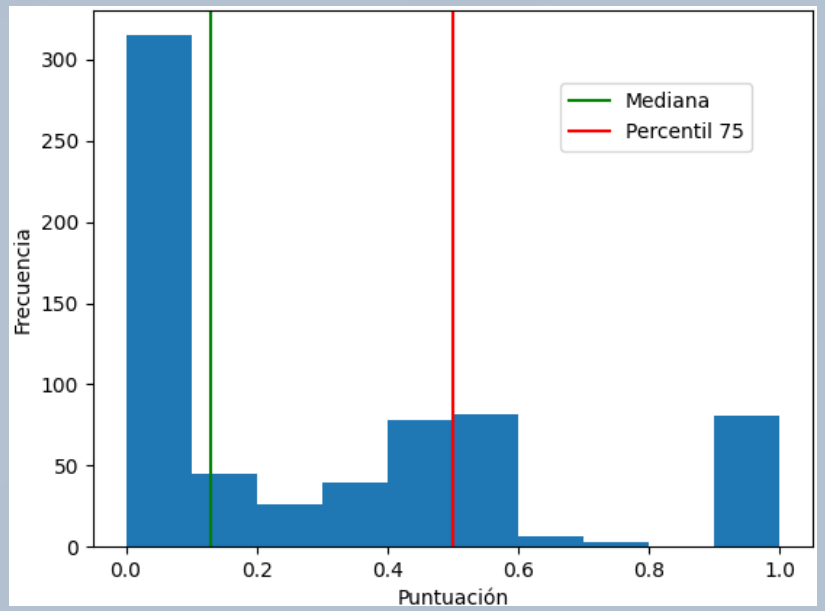
1. Scrapeo de einforma
2. Enlace directo NIF
3. Distancia entre literales de las direcciones

Establecimientos:

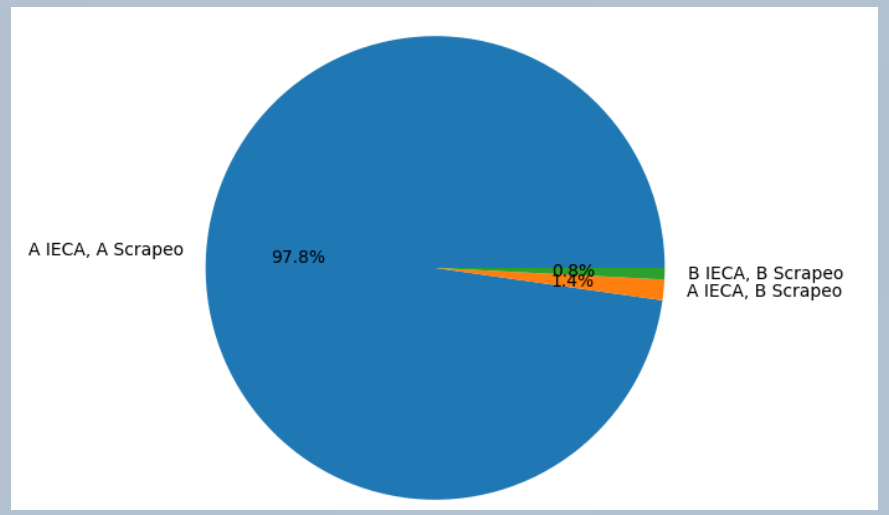
1. Sector bancario
2. Scrapeo páginas estáticas y OpenStreetMap
3. Enlaces:
    1. Direcciones normalizadas
    2. Situación geográfica

Identificación de Empresas y Establecimientos en Andalucía

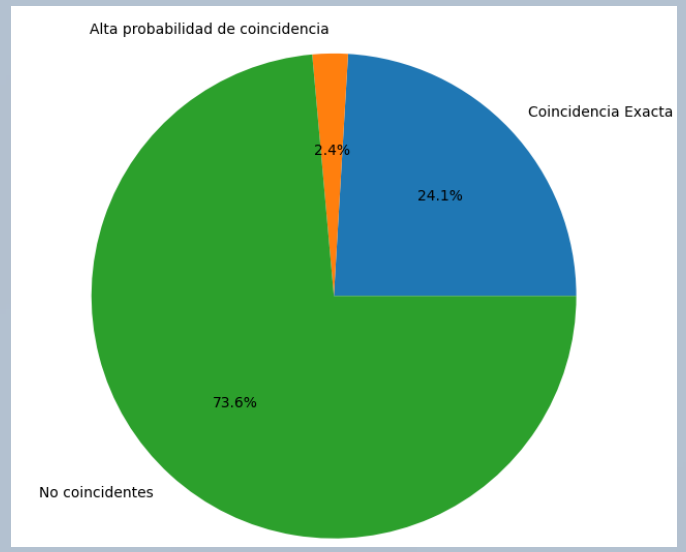
Resultados



Histograma para CNAE 1043
(Fabricación de aceite de oliva)



Altas y Bajas CNAE 1043



Coincidencias IECA/OSM

Buena información de las empresas en einforma:
→ permite la actualización de variables

Establecimientos
→ se necesita más de 1 fuente

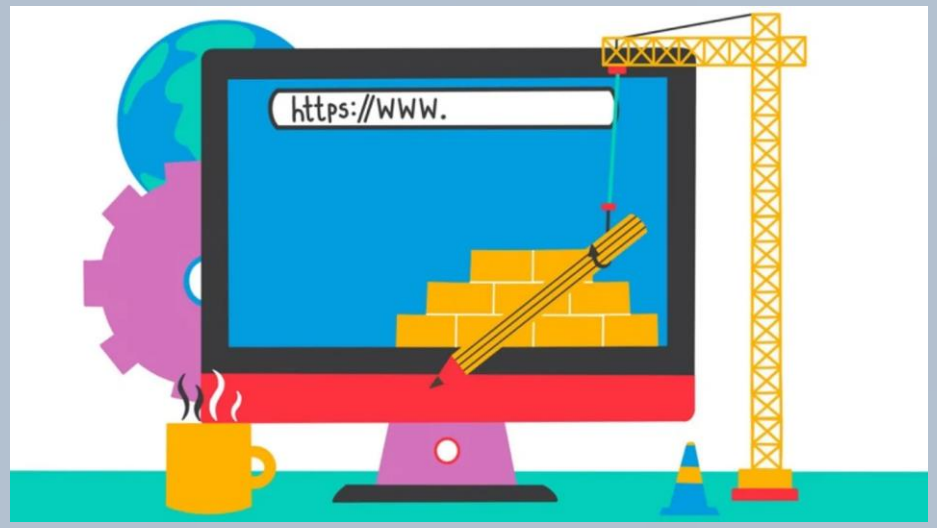
Análisis del carácter innovador de las empresas
Objetivos

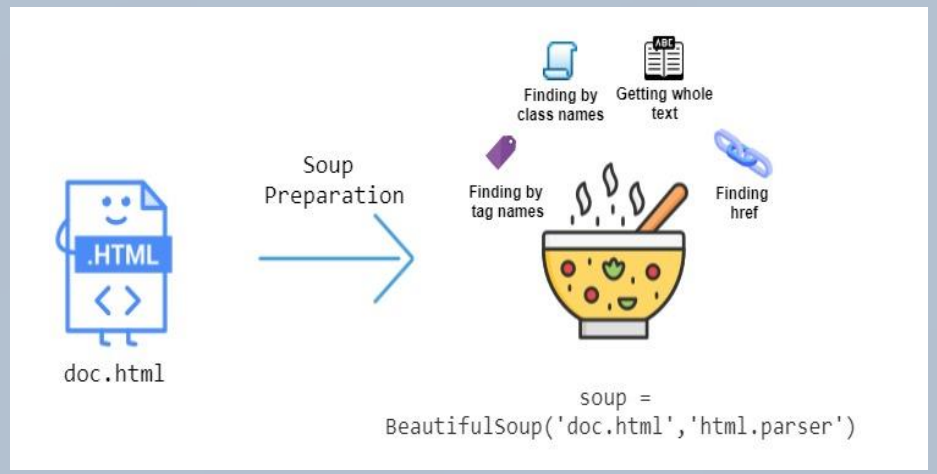
Análisis del carácter innovador de las empresas

1. Modelo de *Machine Learning* de clasificación INNOVACIÓN SÍ/NO
2. Definición de innovación en Andalucía
3. Extrapolación del modelo a empresas de < 10 empleados

Análisis del carácter innovador de las empresas

Metodología. *Web scraping*





- **Obtención de variables descriptivas a partir de la página web con requests y BeautifulSoup**
  - **Meta etiquetas, construcción html, idioma, imágenes**

```
idioma=soup.html['lang']
print(idioma)
```

```
es
```

```
etiquetas_meta = [meta.get('name') for meta in soup.select("meta")]
print(etiquetas_meta)
```

```
[None, 'thumbnail', 'twitter:card', 'twitter:site', 'keywords', 'twitter:image:width',
bileOptimized', 'HandheldFriendly', 'viewport']
```

  - **Texto del cuerpo**

```
pag_parsed=list(soup.stripped_strings)
print(pag_parsed)
```

```
['| Portal Universidad de Sevilla', 'Pasar al contenido principal', 'User account menu', 'Accesibilidad', 'Iniciar sesión'
'Universidad Digital', 'Secretaría virtual', 'Enseñanza virtual', 'Sede electrónica', 'Acceso portafirma', 'Correo', 'Tabl
virtual', 'Cita previa', 'La US', 'Mensaje del Rector', 'Ejes Estratégicos', 'Bienvenidos a la US', 'Plan estratégico', 'R
ings', 'Igualdad', 'Emprendimiento', 'La US en cifras', 'Historia', 'Cultura y Patrimonio', 'Equipo de gobierno', 'Secreta
general', 'Agenda, Documentos y FAQ', 'Órganos Colegiados', 'Órganos dependientes', 'Elecciones', 'Convenios', 'Normativas
'Apertura de cursos', 'Portal de transparencia', 'BOUS', 'Contacto', 'Imagen corporativa', 'Estudiar', 'Estudiar en la US'
'Qué estudiar', 'Grados', 'Másteres', 'Doctorado', 'Dobles titulaciones internacionales', 'Reconocimiento de créditos', 'C
s estudios', 'Acceso y matrícula', 'Becas y ayudas', 'Becas y ayudas', 'Ayudas al estudio', 'Becas asistenciales', 'Becas
yudas de formación', 'Becas y ayudas de movilidad', 'Becas y ayudas culturales y deportivas', 'Premios y distinciones', 'M
lidad de estudiantes', 'Estudiar en Sevilla', 'Estudiantes visitantes', 'Prácticas y empleo', 'Bibliotecas y salas de estu
o', 'Investigar', 'Investigar en la US', 'Investigar', 'Datos de investigación', 'CRAI', 'Espacios de investigación', 'Cen
s mixtos', 'Institutos de investigación', 'Unidad de Cultura Científica', 'OGPI y OPEA', 'FIUS', 'Plan propio', 'Convocato
s', 'Atención al investigador', 'Captación de talento', 'Movilidad de investigadores', 'Doctorado', 'Biblioteca', 'Vivir l
S', 'La US en Sevilla', 'Atención social', 'Cultura', 'Deportes', 'Bibliotecas y salas de estudio', 'Agenda', 'Campus', 'C
ros y departamentos', 'Empresas', 'Prácticas y Empleo', 'Emprendimiento', 'EBC', 'Patentes', 'Cátedras', 'Mecenazgo y cola
aciones', 'Perfil del contratante', 'Internacional', 'US internacional', 'Datos y cifras', 'Alianzas', 'Convenios', 'OGPI'
'Rankings', 'Captación de talento', 'Delegaciones internacionales', 'Oficina Welcome', 'Antes de llegar', 'Vivir en Sevill
```

Análisis del carácter innovador de las empresas

Metodología. Preprocesamiento con minería de texto

```
pag_parsed=list(soup.stripped_strings)
print(pag_parsed)
```

```
['| Portal Universidad de Sevilla', 'Pasar al contenido principal', 'User account menu', 'Accesibilidad', 'Iniciar sesión'
'Universidad Digital', 'Secretaría virtual', 'Enseñanza virtual', 'Sede electrónica', 'Acceso portafirma', 'Correo', 'Tabl
virtual', 'Cita previa', 'La US', 'Mensaje del Rector', 'Ejes Estratégicos', 'Bienvenidos a la US', 'Plan estratégico', 'R
ings', 'Igualdad', 'Emprendimiento', 'La US en cifras', 'Historia', 'Cultura y Patrimonio', 'Equipo de gobierno', 'Secreta
general', 'Agenda, Documentos y FAQ', 'Órganos Colegiados', 'Órganos dependientes', 'Elecciones', 'Convenios', 'Normativas
'Apertura de cursos', 'Portal de transparencia', 'BOUS', 'Contacto', 'Imagen corporativa', 'Estudiar', 'Estudiar en la US'
'Qué estudiar', 'Grados', 'Másteres', 'Doctorado', 'Dobles titulaciones internacionales', 'Reconocimiento de créditos', 'C
s estudios', 'Acceso y matrícula', 'Becas y ayudas', 'Becas y ayudas', 'Ayudas al estudio', 'Becas asistenciales', 'Becas
yudas de formación', 'Becas y ayudas de movilidad', 'Becas y ayudas culturales y deportivas', 'Premios y distinciones', 'M
lidad de estudiantes', 'Estudiar en Sevilla', 'Estudiantes visitantes', 'Prácticas y empleo', 'Bibliotecas y salas de estu
o', 'Investigar', 'Investigar en la US', 'Investigar', 'Datos de investigación', 'CRAI', 'Espacios de investigación', 'Cen
s mixtos', 'Institutos de investigación', 'Unidad de Cultura Científica', 'OGPI y OPEA', 'FIUS', 'Plan propio', 'Convocato
s', 'Atención al investigador', 'Captación de talento', 'Movilidad de investigadores', 'Doctorado', 'Biblioteca', 'Vivir l
S', 'La US en Sevilla', 'Atención social', 'Cultura', 'Deportes', 'Bibliotecas y salas de estudio', 'Agenda', 'Campus', 'C
ros y departamentos', 'Empresas', 'Prácticas y Empleo', 'Emprendimiento', 'EBC', 'Patentes', 'Cátedras', 'Mecenazgo y cola
aciones', 'Perfil del contratante', 'Internacional', 'US internacional', 'Datos y cifras', 'Alianzas', 'Convenios', 'OGPI'
'Rankings', 'Captación de talento', 'Delegaciones internacionales', 'Oficina Welcome', 'Antes de llegar', 'Vivir en Sevill
```


```
def limpieza(lista_palabras):
    stemmer = SnowballStemmer("spanish")
    stop_words = stopwords.words('spanish')
    lista_filtrada = [p.lower() for p in lista_palabras if(p.lower() not in stop_words and p.isalpha())]
    lista_final = [stemmer.stem(p) for p in lista_filtrada]

    return lista_final
```

```
print(limpieza(pag_parsed))
```

```
['accesibil', 'corre', 'rankings', 'iguald', 'emprend', 'histori', 'eleccion', 'conveni', 'normat', 'bous', 'contact', 'estud
i', 'grad', 'master', 'doctor', 'investig', 'investig', 'crai', 'fius', 'convocatori', 'doctor', 'bibliotec', 'cultur', 'depor
t', 'agend', 'campus', 'empres', 'emprend', 'ebc', 'patent', 'catedr', 'internacional', 'alianz', 'conveni', 'ogpi', 'ranking
s', 'cooper', 'profesor', 'pas', 'directori', 'estudi', 'profesor', 'pas', 'alumni', 'search', 'rankings', 'iguald', 'emprend',
'histori', 'eleccion', 'conveni', 'normat', 'bous', 'contact', 'estudi', 'grad', 'master', 'doctor', 'investig', 'investig', 'c
rai', 'fius', 'convocatori', 'doctor', 'bibliotec', 'cultur', 'deport', 'agend', 'campus', 'empres', 'emprend', 'ebc', 'paten
t', 'catedr', 'internacional', 'alianz', 'conveni', 'ogpi', 'rankings', 'cooper', 'profesor', 'pas', 'directori', 'estudi', 'pr
ofesor', 'pas', 'alumni', 'aup', 'estudi', 'actual', 'univers', 'estudi', 'investig', 'cultur', 'deport', 'vist', 'vist', 'vis
t', 'vist', 'vist', 'vist', 'vist', 'vist', 'estudi', 'investig', 'estudi', 'transparent', 'destac', 'estudi', 'investig', 'emp
res', 'directori', 'editorial', 'editorial', 'search', 'encuentran', 'encuentran', 'rankings', 'iguald', 'emprend', 'histori',
'eleccion', 'conveni', 'normat', 'bous', 'contact', 'estudi', 'grad', 'master', 'doctor', 'investig', 'investig', 'crai', 'fiu
s', 'convocatori', 'doctor', 'bibliotec', 'cultur', 'deport', 'agend', 'campus', 'empres', 'emprend', 'ebc', 'patent', 'cated
r', 'internacional', 'alianz', 'conveni', 'ogpi', 'rankings', 'cooper', 'profesor', 'pas', 'directori']
```

- **Limpieza de caracteres (eliminación de símbolos y números)**
- **Eliminación de *stopwords***
- ***Stemming* (obtención de la raíz)**

Análisis del carácter innovador de las empresas
Metodología. Selección y agrupación de variables

- **Formación de clústers de palabras por densidades con sklearn**

Cluster de 'información'

	peso
sostenibil	5,08
ecolog	6,09
medioambient	4,99
infraestr	−0,5
constr	−1,3

- **Selección de variables (etiquetas/palabras) significativas con scipy**

```
Tabla de contingencia para la palabra:  adult
adult        False  True
INNOVACION
0              515     13
1              797      3
p_valor:  0.001104618269509042
```

Figura: Palabra cuya aparición depende de la innovación

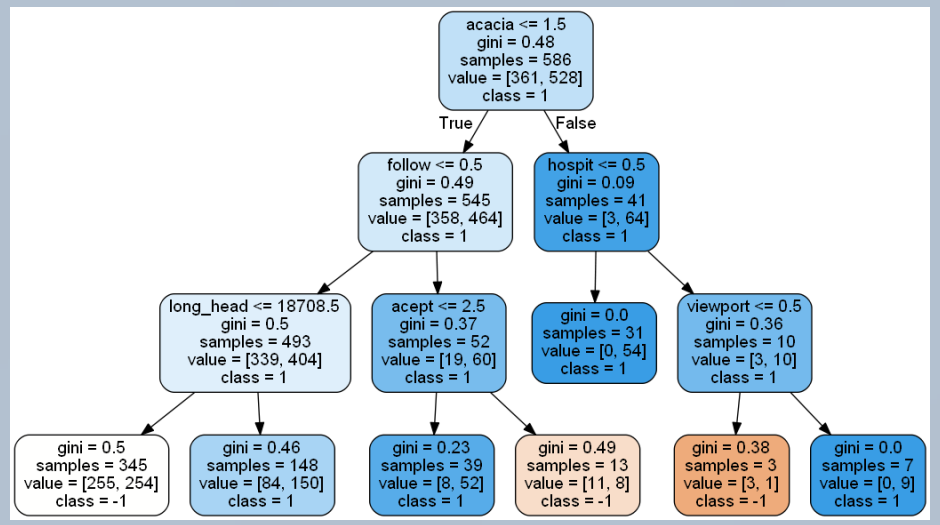
```
Tabla de contingencia para la palabra:  afric
afric        False  True
INNOVACION
0              525      3
1              793      7
p_valor:  0.7482567547446752
```

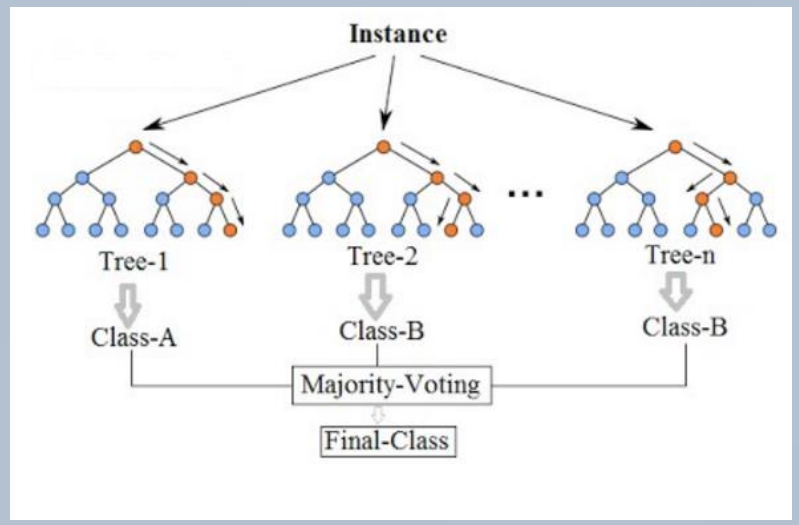
Figura: Palabra cuya aparición es independiente de la innovación

- p-valor no estrictamente pequeño → Eliminación de palabras *noise*
- p-valor 0.05 → Obtención de clusters significativos

Análisis del carácter innovador de las empresas
Metodología. Modelo de clasificación

- **Modelo de clasificación mediante Random Forests con sklearn**



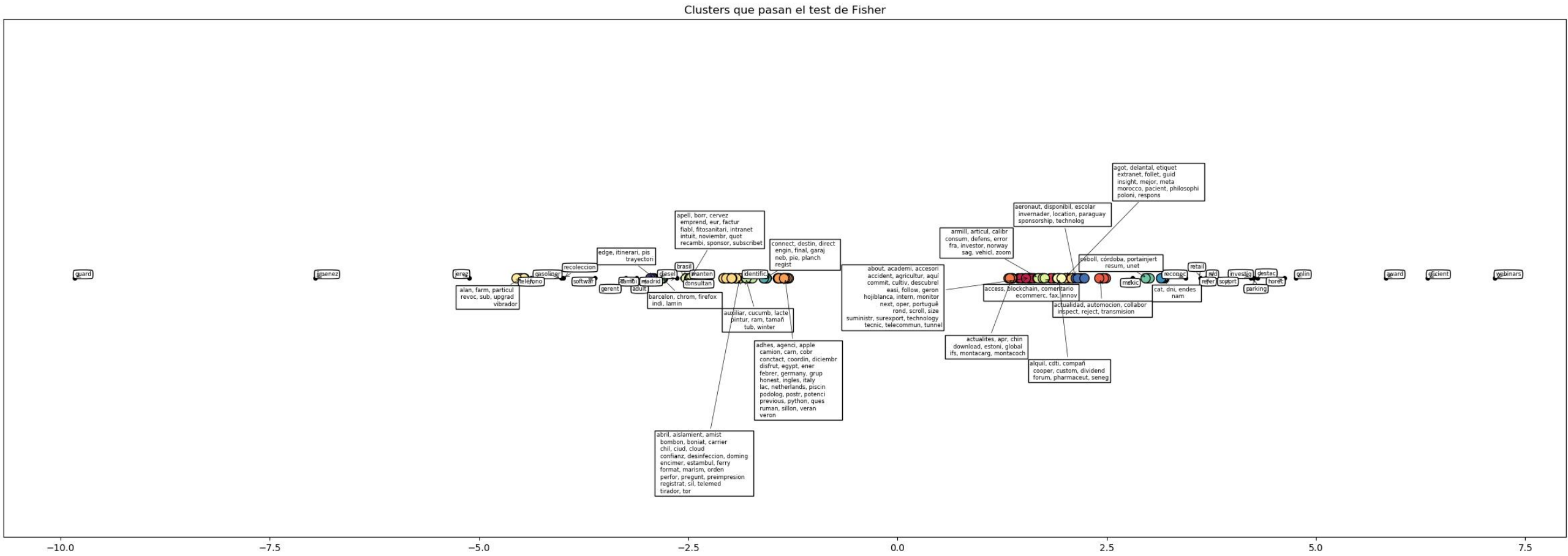


- **Tuneo de hiperparámetros por optimización Bayesia con GPyOpt**

imus instituto de matemáticas universidad de sevilla
Junta de Andalucía
Consejería de Transformación Económica, Industria, Conocimiento y Universidades
Instituto de Estadística y Cartografía de Andalucía

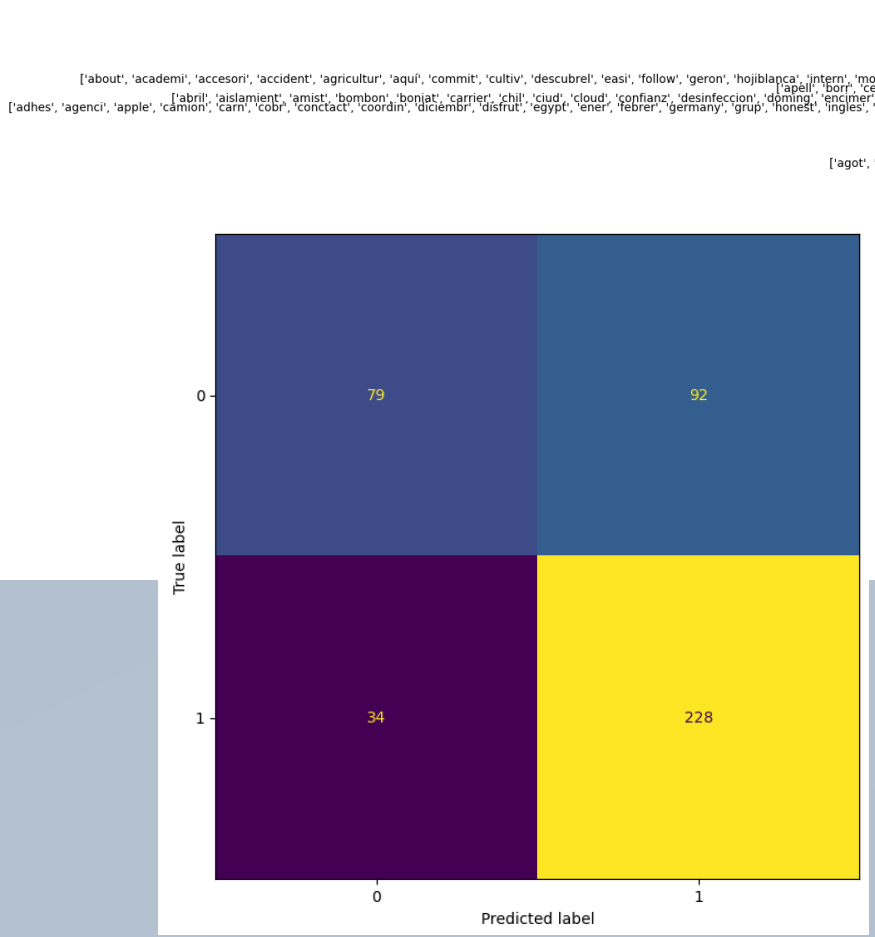
Análisis del carácter innovador de las empresas
Resultados

- **Clústers formados y selección de variables**
  - **Matriz del modelo con 1310 empresas y 4336 variables → 87 variables**



Análisis del carácter innovador de las empresas
Resultados

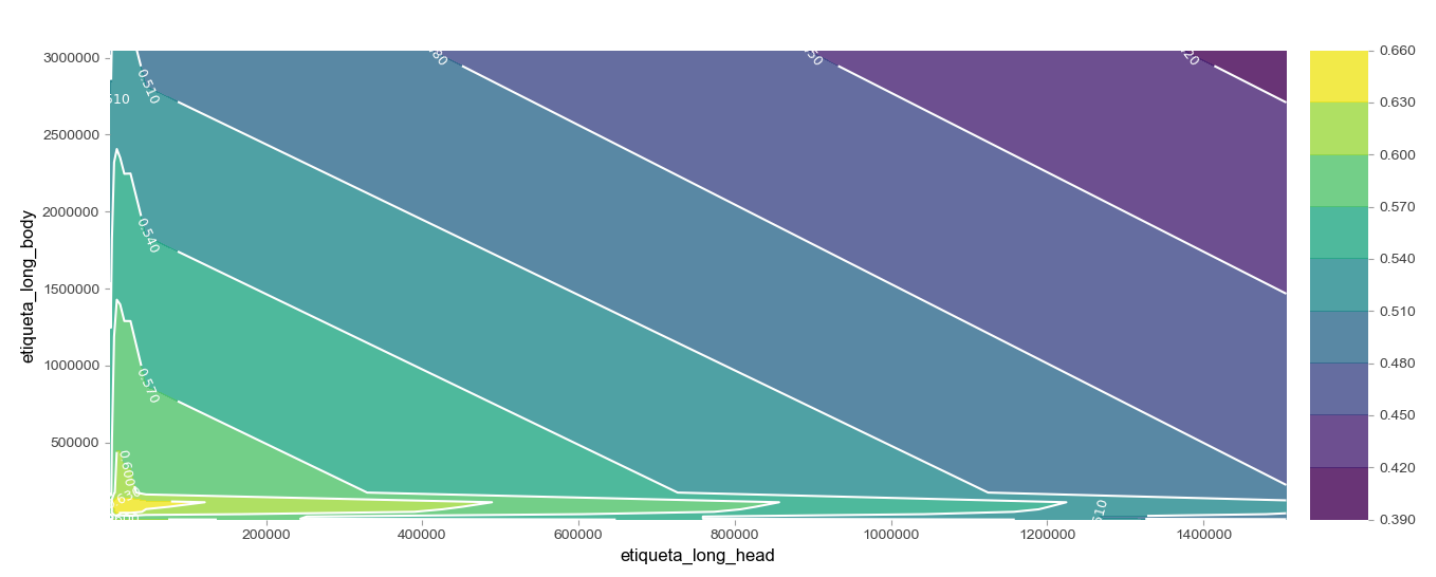
- **Modelo de clasificación**

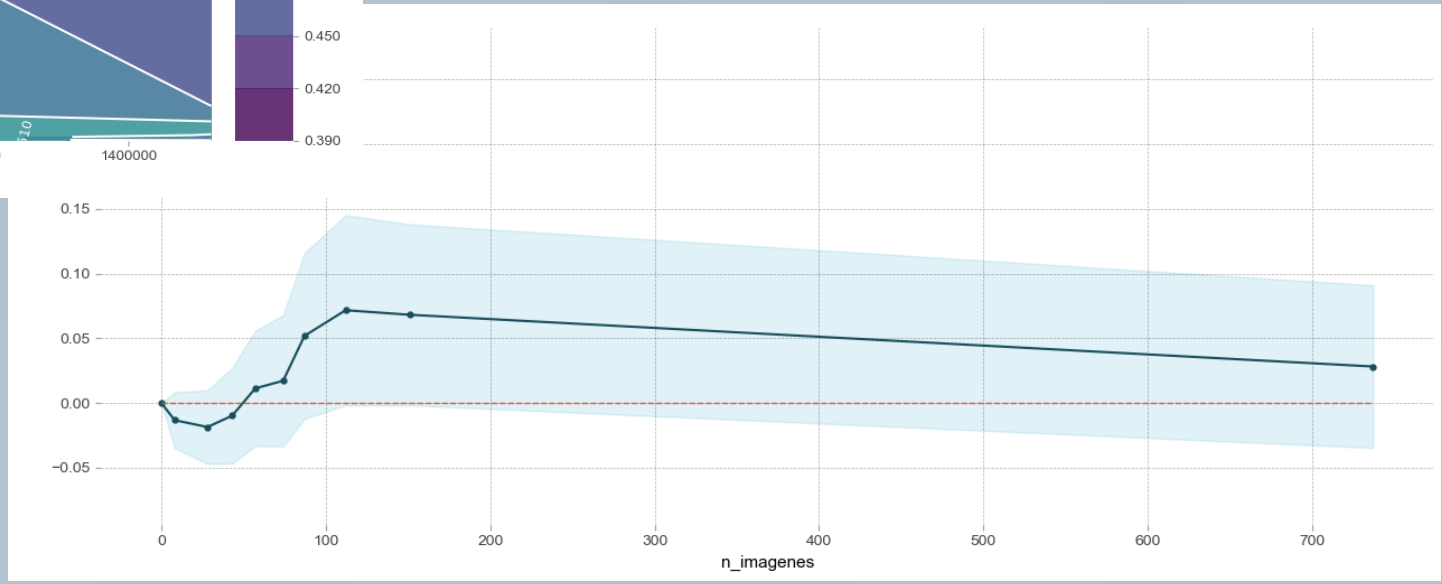


CV Accuracy
70%

Análisis del carácter innovador de las empresas
Resultados

- **Influencias marginales**





Conclusiones

- **Actualización del directorio**
  - ➢ **Buena información de las empresas**
    - ✓ **Actualización de variables (teléfono, página web…)**
    - ✓ **Detección de altas y bajas**
  - ➢ **La presencia de establecimientos es menor: se necesitan varias fuentes**
- **Innovación**
  - ➢ **Definición de innovación más homogénea que la no innovación**
  - ➢ **Importancia de la construcción de la página web**
  - ➢ **Variabilidad del concepto en el tiempo**

Bibliografía

- Muthukadan, B. (2018). Selenium with python. *Retrieved Jan*, *23*, 2019.
- Saurkar, A. V., Pathare, K. G., & Gode, S. A. (2018). An overview on web scraping techniques and tools. *International Journal on Future Revolution in Computer Science & Communication Engineering*, *4(4)*, 363-367.
- Rhodes, B. B., Kim, A. F., & Loomis, B. R. (2015). Vaping the Web: Crowdsourcing and Web Scraping for Establishment Survey Farm Generation. *Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference.*
- Kim, A. E., Loomis, B., Rhodes, B., Eggers, M. E., Liedtke, C., & Porter, L. (2016). Identifying e-cigarette vape stores: description of an online search methodology. *Tobacco Control*, *25(e1)*, e19-e23.
- Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, *14*.
- Daas, P. J., & van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS, (Preprint), 1-13*.

¡Muchas gracias!

Nuria Gómez Vargas, Marina Enguídanos Weyler

ngvargas@us.es

marina.enguidanos@juntadeandalucia.es