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

“Una herramienta de Machine Learning para la actualización y el desarrollo del Directorio de Empresas y Establecimientos con actividad económica en Andalucía” (CEI-23-FQM329)

imus
instituto de matemáticas
universidad de sevilla

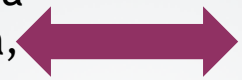
Junta de Andalucía
Consejería de Transformación Económica, Industria, Conocimiento y Universidades
Instituto de Estadística y Cartografía de Andalucía

"Una herramienta de Machine Learning para la actualización y el desarrollo del Directorio de Empresas y Establecimientos con actividad económica en Andalucía" (CEI-23-FQM329)

Elisa Isabel Caballero, Marina Enguidanos, Ana Gema Galera

Rafael Blanquero, Emilio Carrizosa, Nuria Gómez-Vargas, Jasone Ramírez-Ayerbe

Instituto de Estadística y Cartografía de Andalucía (Consejería Transformación Económica, Industria, Conocimiento y Universidades)



Instituto Matemáticas de Universidad de Sevilla (Campus de Excelencia Internacional)

MARCO IMPULSO DESARROLLO DE PROYECTOS SINGULARES DE TRANSFERENCIA EN LOS CEI, RELACIONADOS CON INTELIGENCIA ARTIFICIAL

ESTRATEGIA DE INVESTIGACIÓN E INNOVACIÓN PARA LA ESPECIALIZACIÓN INTELIGENTE DE ANDALUCÍA (RIS3)

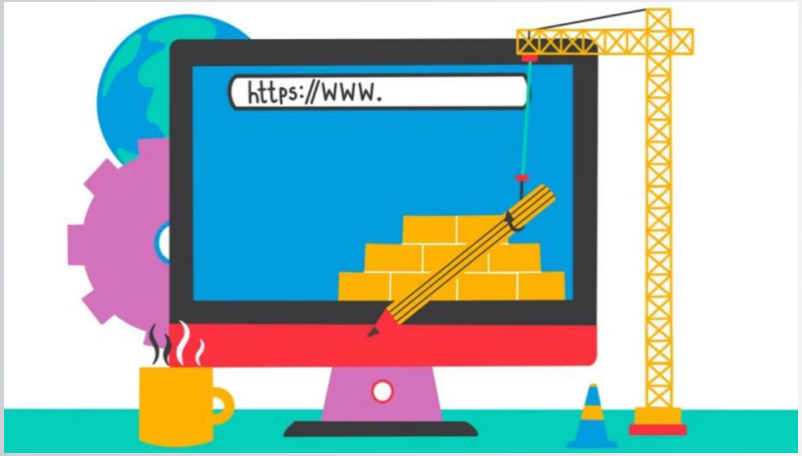
ACTUACIONES COFINANCIADAS POR EL PROGRAMA OPERATIVO FEDER EN ANDALUCÍA 2014-2020

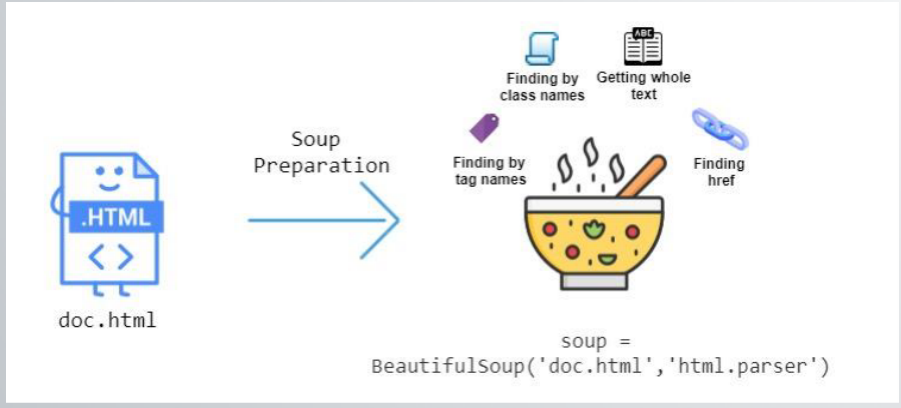
Junta de Andalucía

Unión Europea

Andalucía se mueve con Europa

ANÁLISIS DEL CARÁCTER INNOVADOR DE LAS EMPRESAS
METODOLOGÍA. *WEB SCRAPING*





- *Obtención de variables descriptivas a partir de la página web*

```
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', 'F
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', 'Cer
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', 'F
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', 'Cer
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 de puntuación, números…)*
- *Eliminación de stopwords*
- *Stemming (obtención de la raíz)*

ANÁLISIS DEL CARÁCTER INNOVADOR DE LAS EMPRESAS
METODOLOGÍA. *SELECCIÓN Y AGRUPACIÓN DE VARIABLES*

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

```
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
```

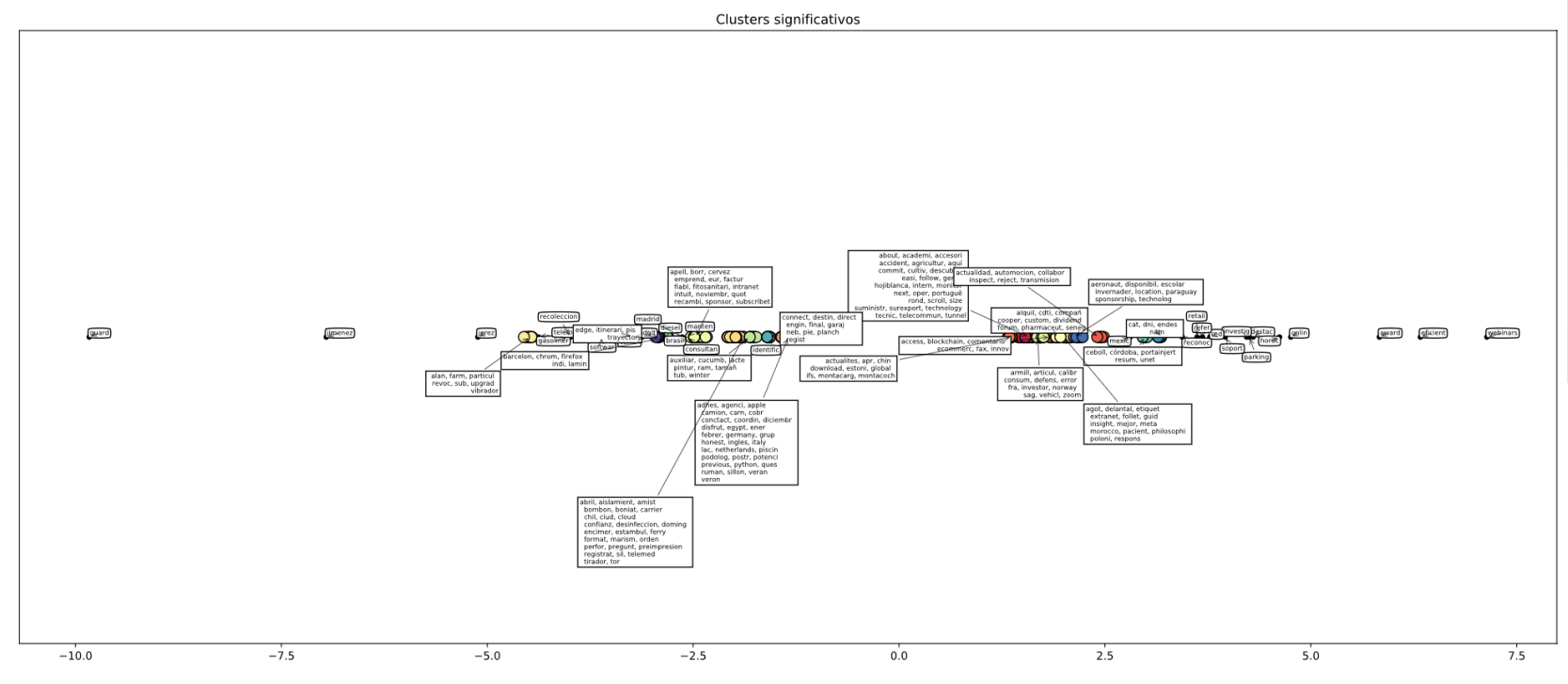
Figura: Palabra cuya aparición es independiente de la innovación

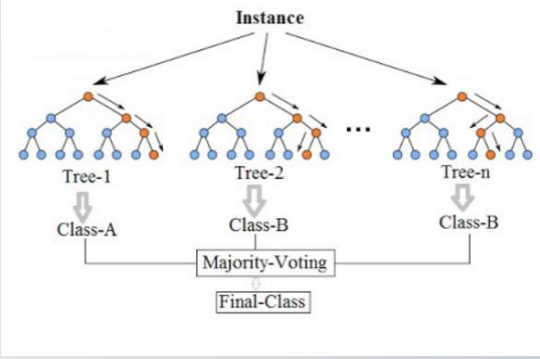
- p-valor no estrictamente pequeño $\rightarrow$ Eliminación de palabras *noise*
- p-valor $\leq$ 0.05 $\rightarrow$ Obtención de clusters significativos

ANÁLISIS DEL CARÁCTER INNOVADOR DE LAS EMPRESAS
METODOLOGÍA. *SELECCIÓN Y AGRUPACIÓN DE VARIABLES*

- *Formación de **clústers de palabras***

ANÁLISIS DEL CARÁCTER INNOVADOR DE LAS EMPRESAS
RESULTADOS

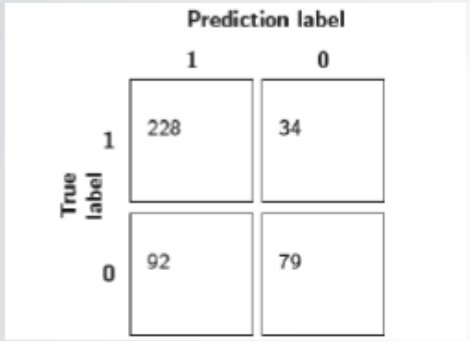




Variables del modelo de Random Forests

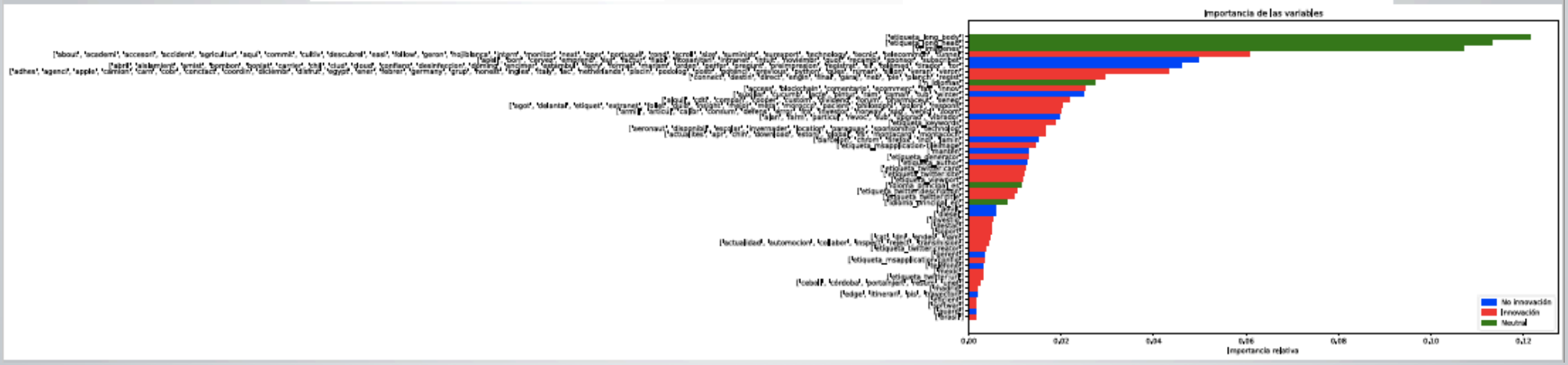
- 1310 instancias (empresas).
- ~~4336~~ 87 variables:
  - ~~4204 palabras~~ 48 clusters (30 outliers + 18 grupos).
  - ~~132~~ 39 etiquetas html.
- Respuesta: Innovadora SÍ/NO.

ANÁLISIS DEL CARÁCTER INNOVADOR DE LAS EMPRESAS
RESULTADOS



Medidas de efectividad

- 0.7 precisón
- 0.87 sensibilidad
- 0.46 especificidad



CONCLUSIONES

- *Definición de innovación: más homogénea que la no innovación*
- *Importancia de la construcción de la página web*
- *Variabilidad del concepto en el tiempo usando la importancia de las variables*
- *Por hacer:*
  - *Extrapolación a pequeñas empresas*
  - *Mejora de la interpretabilidad*