Machine Learning defines innovation

Rafael Blanquero Bravo[1], Elisa Isabel Caballero Ruiz[2], Emilio Carrizosa Priego[1], Marina Enguidanos Weyler[2], Ana Gema Galera Pozo[2], **Nuria Gómez-Vargas**[1], Jasone Ramírez-Ayerbe[1]

[1]Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Seville, Spain
[2]Instituto de Estadística y Cartografía de Andalucía (IECA), Seville, Spain

New Bridges between Mathematics and Data Science
Minisymposia: Mathematical Optimization for Data-Driven Decision-Making
Nov 8^{th} – 11^{th} 2021, Valladolid, Spain

Consejería de Transformación Económica, Industria, Conocimiento y Universidades

Instituto de Estadística y Cartografía de Andalucía

Content

1. Approaching the problem
2. Text Mining
   - Web Scraping
   - Preprocessing
3. Grouping and feature selection techniques
4. Classification model
   - Random Forest
   - Interpretation of results and definition of innovation
5. Conclusions and further work

Approaching the problem

Characterizing a company

(a) Official surveys, General Business Register, Chamber of Commerce, etc.

(b) Data providing companies

Figure: Traditional methodology

Approaching the problem

Characterizing a company

(a) Official surveys, General Business Register, Chamber of Commerce, etc.

(b) Data providing companies

Figure: Traditional methodology

Proposed new methodology: Web scraping [ten Bosch et al., 2018]

- **Field of growing importance** for Official Statistics Institutes.
- Valuable way to **supplement** administrative sources and metadata systems.
- Web data are more **volatile** and usually **unstructured** but in many cases also **richer** and **more frequently updated**.

Content

1. Approaching the problem
2. Text Mining
   - Web Scraping
   - Preprocessing
3. Grouping and feature selection techniques
4. Classification model
   - Random Forest
   - Interpretation of results and definition of innovation
5. Conclusions and further work

State-of-the-Art

- Linkage between business websites addresses (URLs) and a business population frame [Van Delden et al., 2019].
- Supervised Learning model to determine if a company is innovative by studying the text on its website [Daas and van der Doef, 2020].

State-of-the-Art

- Linkage between business websites addresses (URLs) and a business population frame [Van Delden et al., 2019].
- Supervised Learning model to determine if a company is innovative by studying the text on its website [Daas and van der Doef, 2020].

State-of-the-Art

- Linkage between business websites addresses (URLs) and a business population frame [Van Delden et al., 2019].
- Supervised Learning model to determine if a company is innovative by studying the text on its website [Daas and van der Doef, 2020].

(a) Text in html body

(b) Html tags

Figure: Scraped variables

Text Mining

Preprocessing

- Removal of characters: numbers and punctuation marks.
- Converting words to lower case.
- Words cleaning by language detection:
  - Detection and removal of stop words.
  - Stemming (mapping of the different morphological variants to their base form).

```
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']
```

Figure: Processed text

Content

1. Approaching the problem
2. Text Mining
   - Web Scraping
   - Preprocessing
3. Grouping and feature selection techniques
4. Classification model
   - Random Forest
   - Interpretation of results and definition of innovation
5. Conclusions and further work

Grouping and feature selection techniques

Clustering

- Grouping words by their weights β_j in a classification model.
  - Penalized logistic regression: ridge regression

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \cdot \sum_{j=1}^{p} \beta_j^2 \quad , \lambda > 0$$

- Density-based clustering: DBSCAN [Ester et al., 1996].
  - Designed for the clustering of large noisy datasets.
  - Clusters with arbitrary shape.
  - Detection of outliers.

Grouping and feature selection techniques

Clustering

- Grouping words by their weights β_j in a classification model.
  - Penalized logistic regression: ridge regression

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \cdot \sum_{j=1}^{p} \beta_j^2 \quad , \lambda > 0$$

- Density-based clustering: DBSCAN [Ester et al., 1996].
  - Designed for the clustering of large noisy datasets.
  - Clusters with arbitrary shape.
  - Detection of outliers.

Grouping and feature selection techniques

Clustering

- Grouping words by their weights β_j in a classification model.
  - Penalized logistic regression: ridge regression

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \cdot \sum_{j=1}^{p} \beta_j^2 \quad , \lambda > 0$$

- Density-based clustering: DBSCAN [Ester et al., 1996].
  - Designed for the clustering of large noisy datasets.
  - Clusters with arbitrary shape.
  - Detection of outliers.

Clustering

- Grouping words by their weights β_j in a classification model.
  - Penalized logistic regression: ridge regression

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \cdot \sum_{j=1}^{p} \beta_j^2 \quad , \lambda > 0$$

- Density-based clustering: DBSCAN [Ester et al., 1996].
  - Designed for the clustering of large noisy datasets.
  - Clusters with arbitrary shape.
  - Detection of outliers.

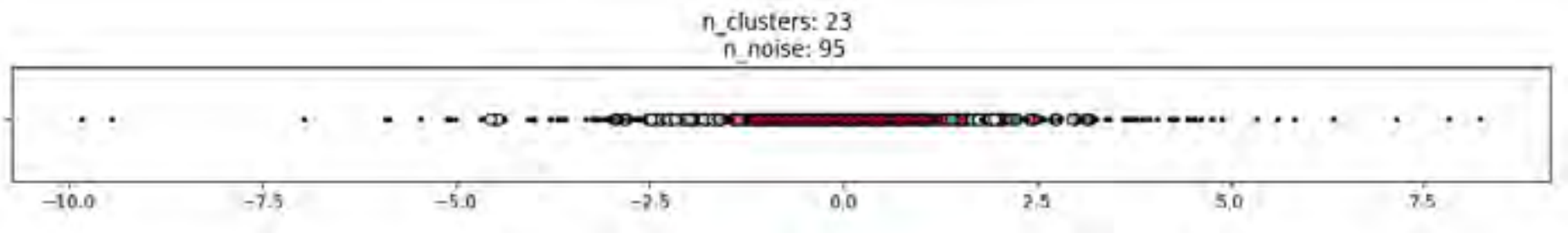

Figure: Clustering according to the weights

Fisher's Exact Test

Determining the dependence between two categorical variables: X: the variable appears yes/no in the company's website; Y: the company is innovative yes/no.

	Appearance	Non-appearance
Innovative	a	b
Non-innovative	c	d

Table: Contingency table

p-value $\leq$ threshold $\rightarrow$ we assume that there is a significant dependence between the appearance and the innovative character.

Fisher's Exact Test

Determining the dependence between two categorical variables: X: the variable appears yes/no in the company's website; Y: the company is innovative yes/no.

	Appearance	Non-appearance
Innovative	a	b
Non-innovative	c	d

Table: Contingency table

p-value $\leq$ threshold $\rightarrow$ we assume that there is a significant dependence between the appearance and the innovative character.

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

(a) Rejecting independence

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

(b) Assuming independence

Figure: Examples of Fisher's Test

Grouping and feature selection techniques

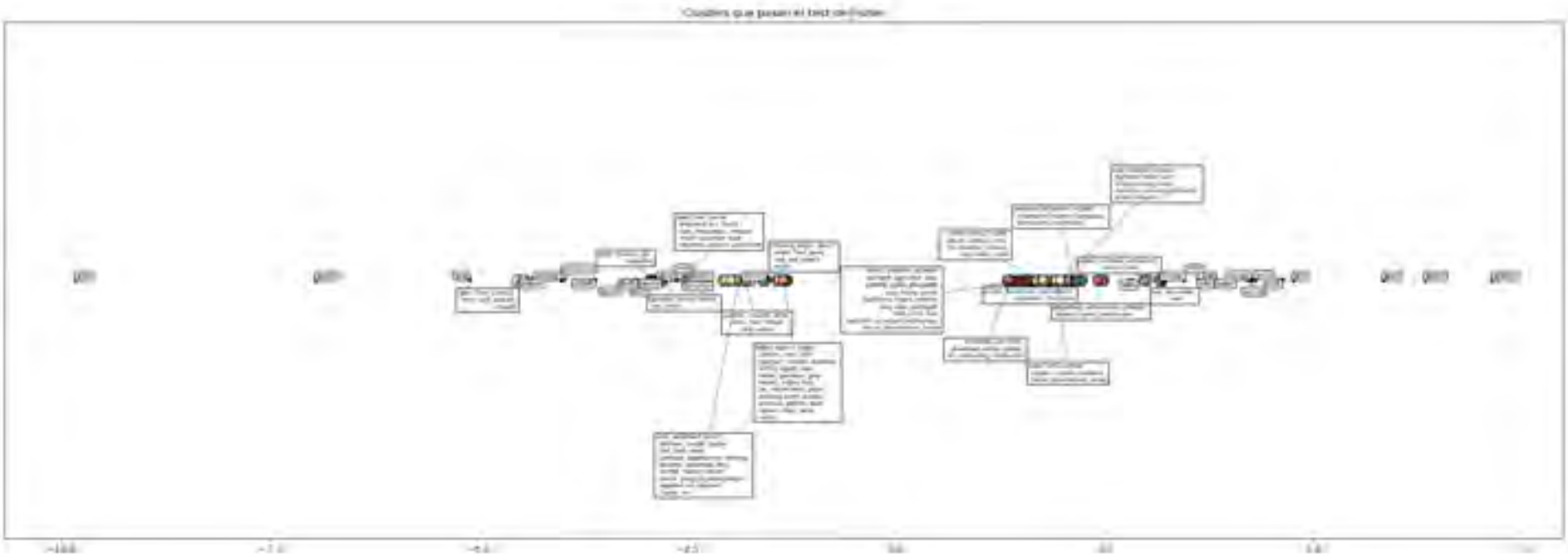

Figure: Selected clusters for the features of the model

Content

1. Approaching the problem
2. Text Mining
   - Web Scraping
   - Preprocessing
3. Grouping and feature selection techniques
4. Classification model
   - Random Forest
   - Interpretation of results and definition of innovation
5. Conclusions and further work

Classification model

Random Forest

Model variables

- 1310 instances (companies).
- ~~4336 features~~ 87 features:
  - ~~4204 words~~ 48 clusters (30 noise + 18 groups).
  - ~~132 html tags~~ 39 html tags.
- Response: Innovative YES/NO.

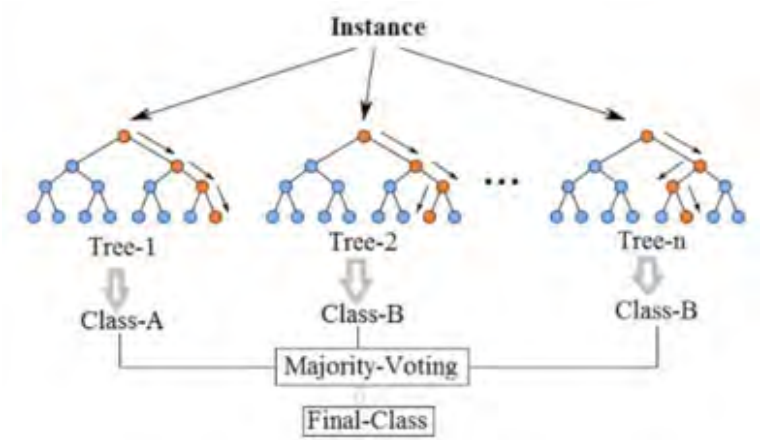

Figure: Machine Learning model: Random Forest

Classification model

Random Forest

Model variables

- 1310 instances (companies).
- ~~4336 features~~ 87 features:
  - ~~4204 words~~ 48 clusters (30 noise + 18 groups).
  - ~~132 html tags~~ 39 html tags.
- Response: Innovative YES/NO.

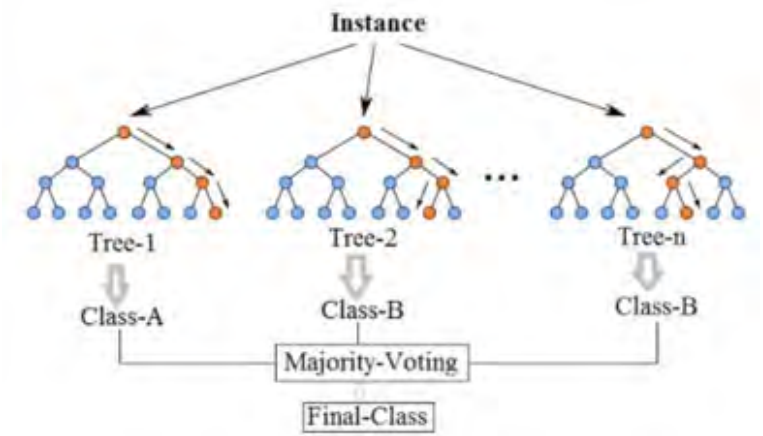

Figure: Machine Learning model: Random Forest

		Prediction label 1	Prediction label 0
True label	1	228	34
	0	92	79

Table: Confusion matrix over test set

Effectiveness measures

- 0.7 CV accuracy
- 0.87 sensitivity
- 0.46 specificity

Classification model

Interpretation of results and definition of innovation

Figure: Variable relative importance

Classification model

Interpretation of results and definition of innovation

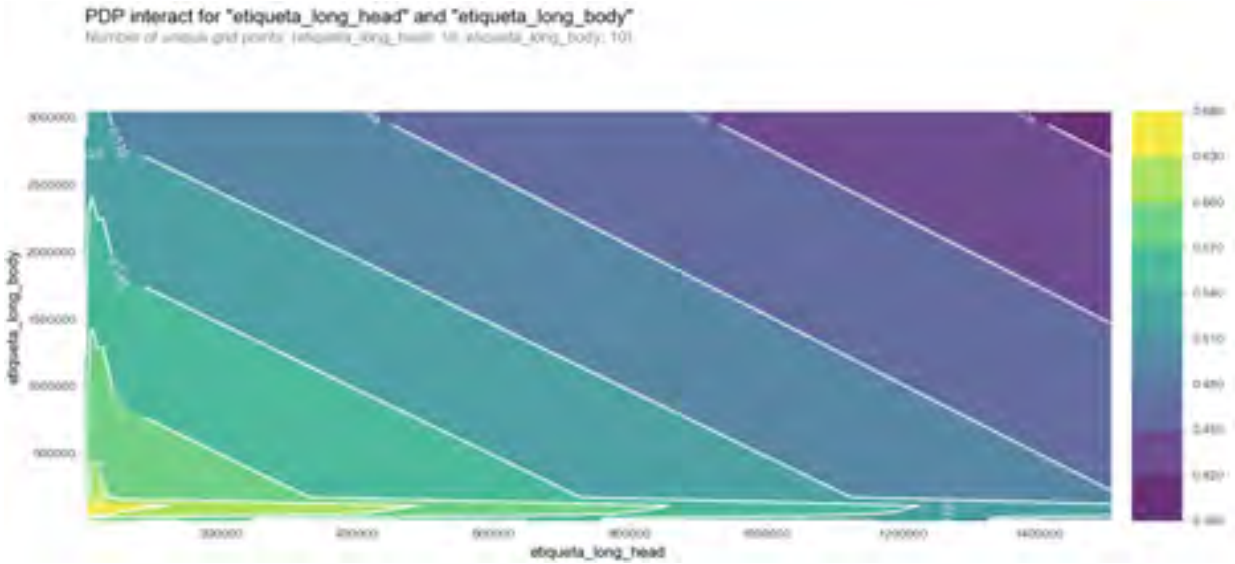

Figure: Partial dependence plot (PDP): *long head* and *long_body*

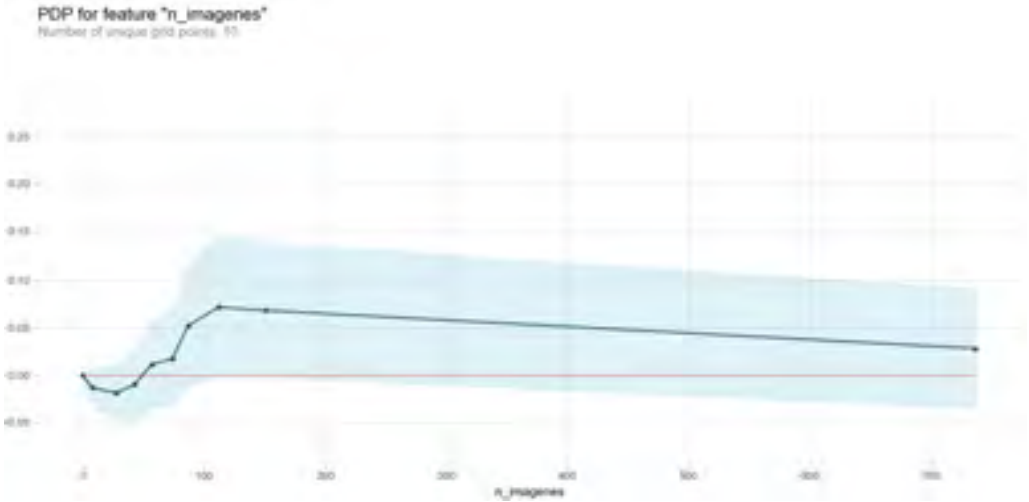

Figure: Individual Conditional Expectation (ICE) plot: *n_imagenes*

Content

1. Approaching the problem
2. Text Mining
   - Web Scraping
   - Preprocessing
3. Grouping and feature selection techniques
4. Classification model
   - Random Forest
   - Interpretation of results and definition of innovation
5. Conclusions and further work

Conclusions and further work

Conclusions

- The webpage content of innovative companies in Andalusia is more homogeneous than that of non-innovative ones.
- How the webpage is constructed prevails over its content.
- It is possible to carry out a study in time of the concept of innovation through the importance of the variables.

Conclusions and further work

Conclusions

- The webpage content of innovative companies in Andalusia is more homogeneous than that of non-innovative ones.
- How the webpage is constructed prevails over its content.
- It is possible to carry out a study in time of the concept of innovation through the importance of the variables.

Further work

- Extrapolate the model to small companies.
- Enhance interpretability:
  - Exploiting background knowledge, e.g., semantic meaning clustering: Must-Link and Cannot-Link constraints (C-DBSCAN [Ruiz et al., 2010]).

References

Piet JH Daas and Suzanne van der Doef. Detecting innovative companies via their website. *Statistical Journal of the IAOS*, 2020.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 1996.

Carlos Ruiz, Myra Spiliopoulou, and Ernestina Menasalvas. Density-based semi-supervised clustering. *Data mining and knowledge discovery*, 21(3), 2010.

Olav ten Bosch, Dick Windmeijer, Arnout van Delden, and Guido van den Heuvel. Web scraping meets survey design: combining forces. In *Big Data Meets Survey Science Conference, Barcelona, Spain*, 2018.

Arnout Van Delden, Dick Windmeijer, and Olav Ten Bosch. Finding enterprise websites. In *European Establishment Statistics Workshop, Bilbao, Spain*, 2019.

Thank you for your attention!

ngvargas@us.es

This work has been carried out within the framework of the research project "Una herramienta de Machine Learning para la actualización y el desarrollo del Directorio de Empresas y Establecimientos con actividad en Andalucía" CEI-23-FQM329, which has received funding from the Junta de Andalucía (Consejería de Economía, Conocimiento, Empresas y Universidad).