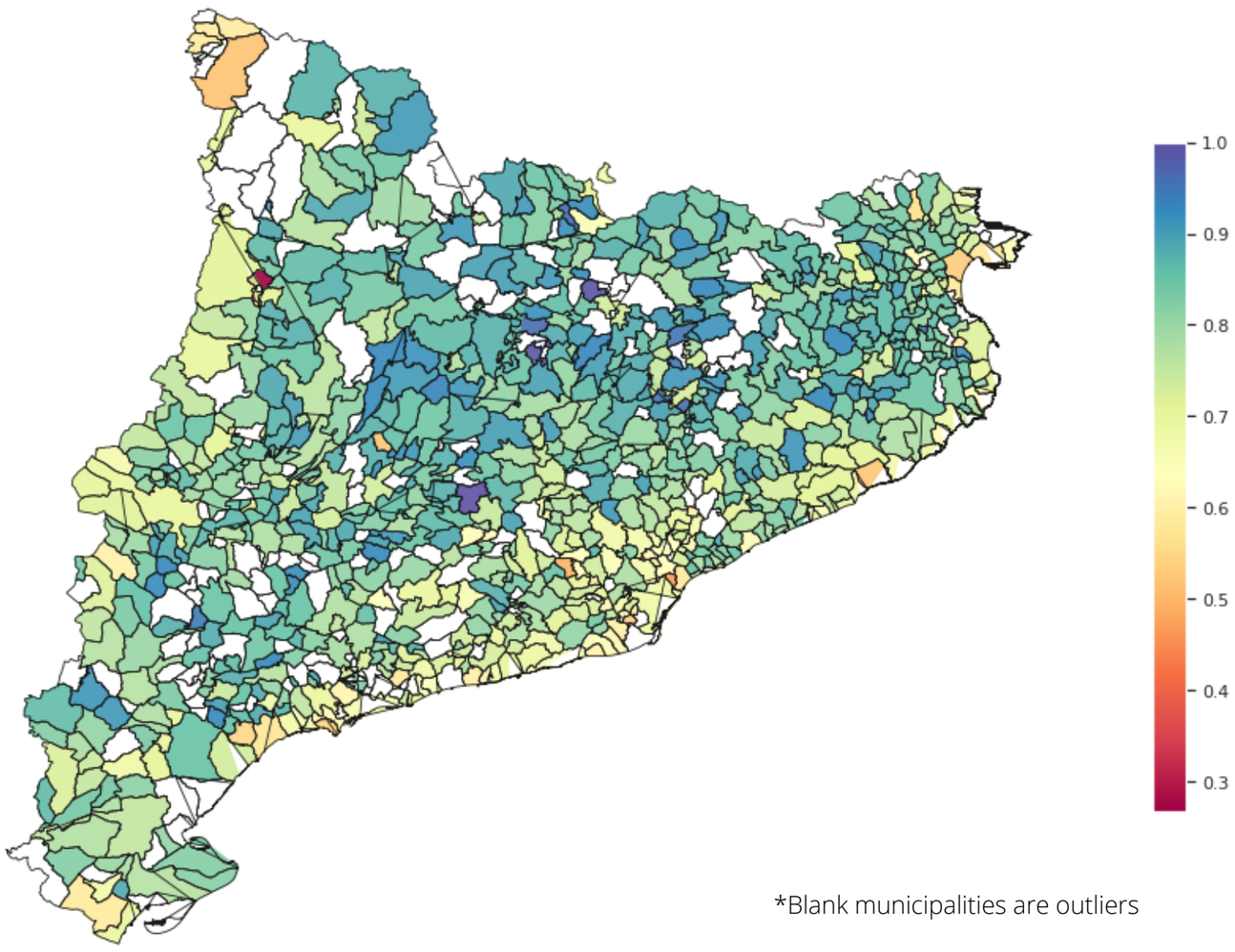
NATIVE SPEAKERS IN CATALAN MUNICIPALITIES

Enric Reverter & Pim Schoolkate

PROBLEM

Catalan language has experienced difficulties during the last century, especially during the persecution after the Spanish Civil War. As such, Catalan speakers have been oscillating in numbers throughout generations. The aim of this project is to analyze the number of speakers taking into account contextual variables such as the level of education attained or the number of immigrants at a municipality level. To do so, data from Idescat is used.

Distribution of Catalan Speakers by municipality



*Blank municipalities are outliers

OUTCOME

The best model by itself was found to be a SVM. However, a soft voting ensemble consisting of the LDA, QDA, KNN, LR, and SGD outperforms. This classifier reached a weighted F1-Score of 68.6%. Moreover, the model generalizes well since the F1-score on the test set is 69.6%.

The most important predictors were extracted from the RF. The percentage of Catalan speakers is largely affected by elderly population, elections turnout, land ownership and number of dwellings, higher level of education, nationality of population, registered unemployment, and tourists accommodations. Results seem to indicate that more rural municipalities have a larger share of Catalan speakers, as depicted by the distribution of Catalan speakers by municipality.

A major limitation was the different times of measurement of the indicators. These should have been scaled with the population during the time of measurement, not with the current population.

Unexpectedly, nearly all models achieved the same F1-score, hinting towards some problem with the approach or data. The above limitation or some other unobserved pattern in the data might be of cause. Given that the model generalized well, this was not seen as a problem.

METHOD

Data description

For each municipality, the data provided indicators considering culture, economics, elections, the environment, labour, population, quality of life and territory. **Catalan speech** is expressed as a percentage of the population.

Preprocessing

The **target** variable was transformed into a 3 categorical variable, namely low (Q1), middle (Q2-Q3), and high (Q4)

Scaling was applied as most indicators were reported in terms of absolute amount of population.

Imputation methods were used to account for the large amount of **missing values** in some indicators. K-nearest neighbours was found to yield the best result, as confirmed by the Kolmogorov-Smirnov tests. Indicators which consisted for 30% of missing values were removed.

A multivariate model-based approach named Isolation Forest was used to detect 3% of the data as **outliers.**

The high correlations found between variables required to apply Principal Components Analysis and **reduce dimensionality**, while also describing categorical as numerical variables.

Then, clustering is explored using two methods, K-Means, where the number of **clusters** need to be predefined, and Affinity Propagation, which directly computed the number of clusters.

Validation and Sampling Strategy

Data is **split** into stratified train (77%) and test (33%) sets. Leave-one-out cross-validation is used during the model training with F1-score as the comparison measure.

Modeling Strategy

10 Models were explored, consisting of Linear Discriminant Analysis, Quadratic Discriminant Analysis, Gaussian Naive Bayes, Linear Regression, Stochastic Gradient Descent, K-Nearest Neighbours, Decision Tree, Random Forest, Support Vector Machine, and a neural network structure. Random search walks were used in order to find the optimal set of parameters for each of the above mentioned models. The best models were selected and combined into a **Voting Classifier**.