**Origin of data**

Site Url:

* The crime count categorized by Region, Crime Type, and Year (spanning from 2013 to 2022
* The population percentage in thousands is categorized by Gender, Region, Age, and Year (covering the years 2013 to 2023)
* The poverty percentage categorized by Region and Year (covering the years 2013 to 2023)
* The education percentage categorised by class of education, Year( covering 2023 to 2014 in quarters), Sex and Region
* The no of crimes categorized by Sex, Region and year(2013 to 2022).



**Cleaning and Transformation**

1. Converted the provided dataset from an unstructured format into a structured, tabular form using Python, facilitating easier data manipulation and analysis.

2. Eliminated any trailing spaces present in the data to ensure consistency and accuracy in text-based fields.

3. Excluded total count information from the analysis, as it was determined to be extraneous and potentially detrimental to the efficiency of machine learning algorithms.

4) Join the formatted data origin data specified In the section above to create the required final datatset for analysis

The main fields included in the final dataset are

1. Region( 19 regions in total)
2. Percentage of Crimes created based on Total crimes and population percentage(perc\_delito)
3. Percentage of poverty (perc\_pobreza)
4. Percentage of Gender segregated by male and female(perc\_hombres,perc\_mujeres)
5. Percentage of education across different education classes(Analfabetos,Educación\_primaria,Educación\_superior,Escuela\_Profesional,Escuela\_Secundaria,Estudios\_primarios\_incompletos,Licenciado)
6. Year( 2013 to 2022)

In the case of education percentage categorized by classes, a separate transformation was required to do backfilling of data for 2013 based on 2014 as this data was missing .

The final dataset after transformationis provided below



**Exploratory Analysis**

1. Correlation between region and sex

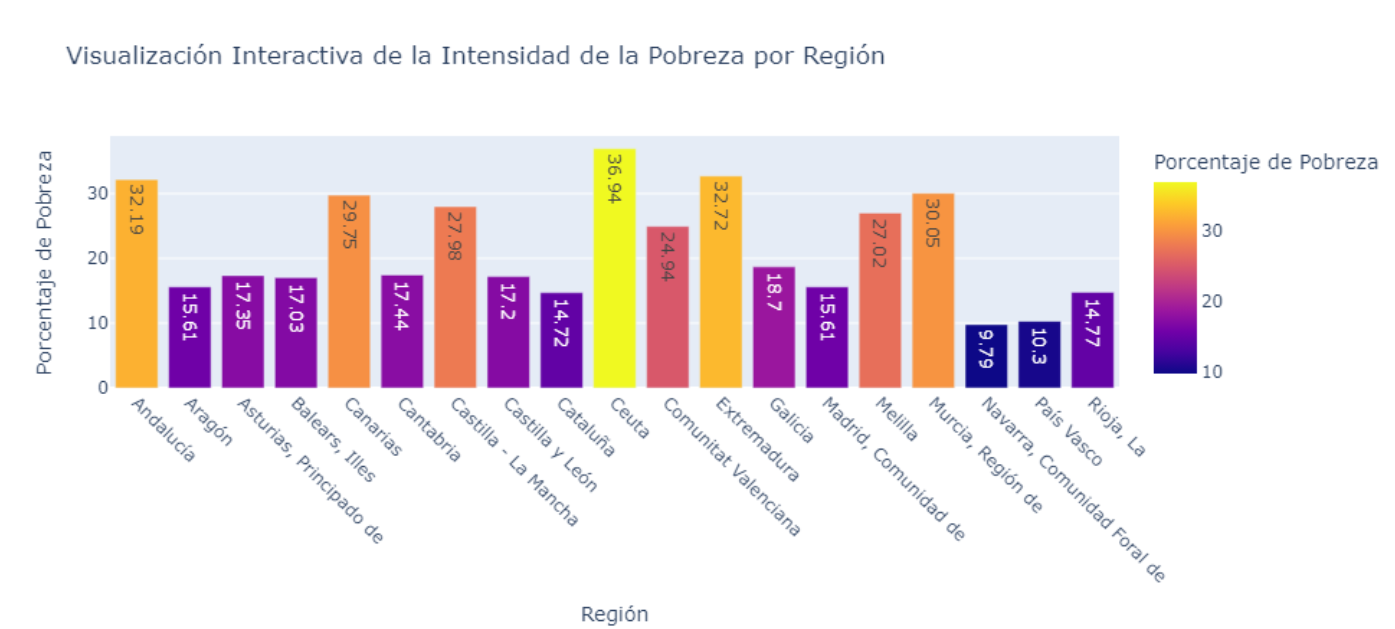
Below radar plot shows the relationship between Region and gender in relation of the mean crime percentage. Even though there is a difference across the regions for gender, this variation is constant across the regions and usage of this feature for modelling doesnt look promising.

A blue circle with white text

Description automatically generated

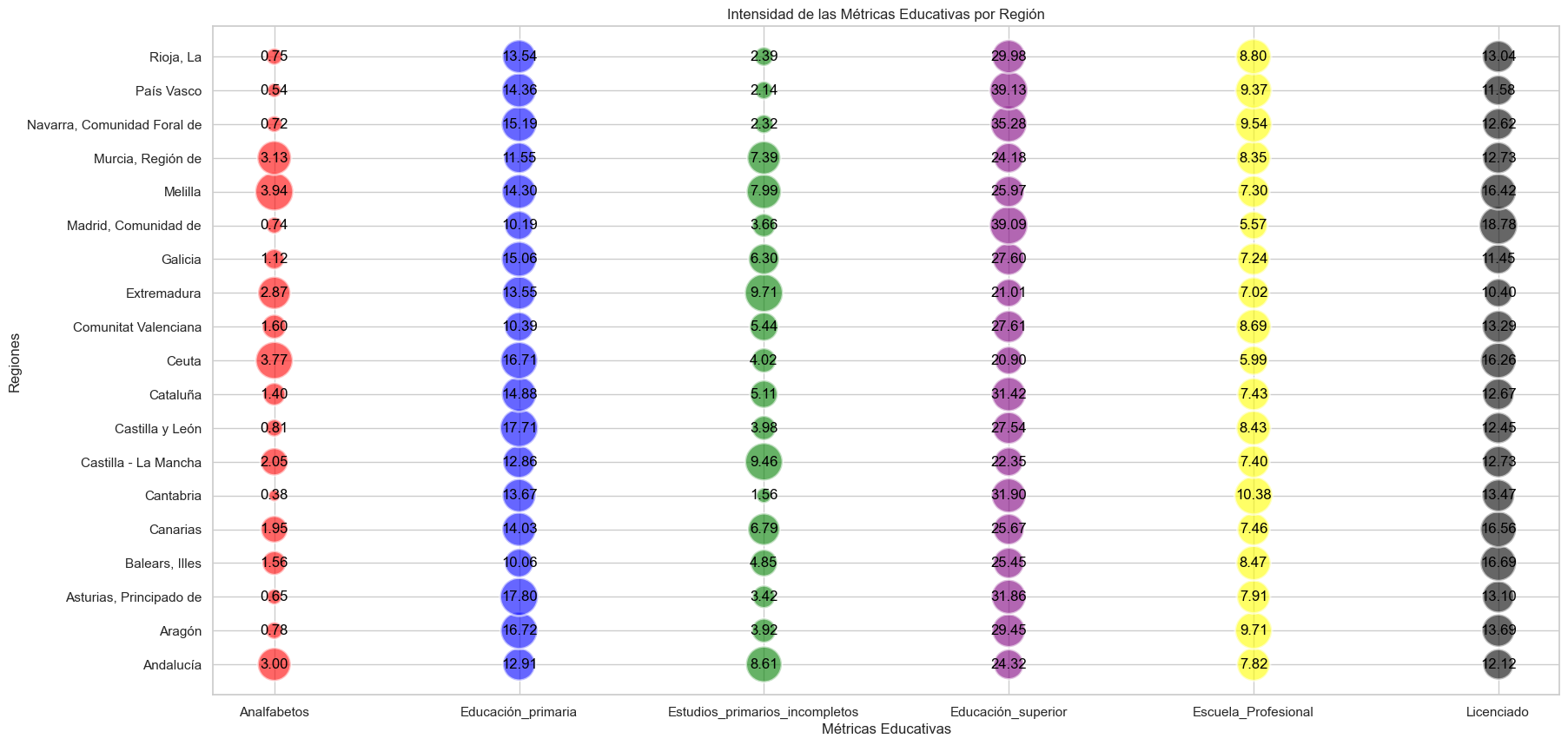
1. Correlation between region and poverty percentage

From the graph, we can see clear pattern difference across regions based on poverty percentage. This gives an initial hint that poverty percentage can be featiure that can be used for the modelling step. Country Ceuta is the region with the highest poverty percentage and Navarra,Region de is the one with the lowest poverty percentage.



Correlation between Education and Region

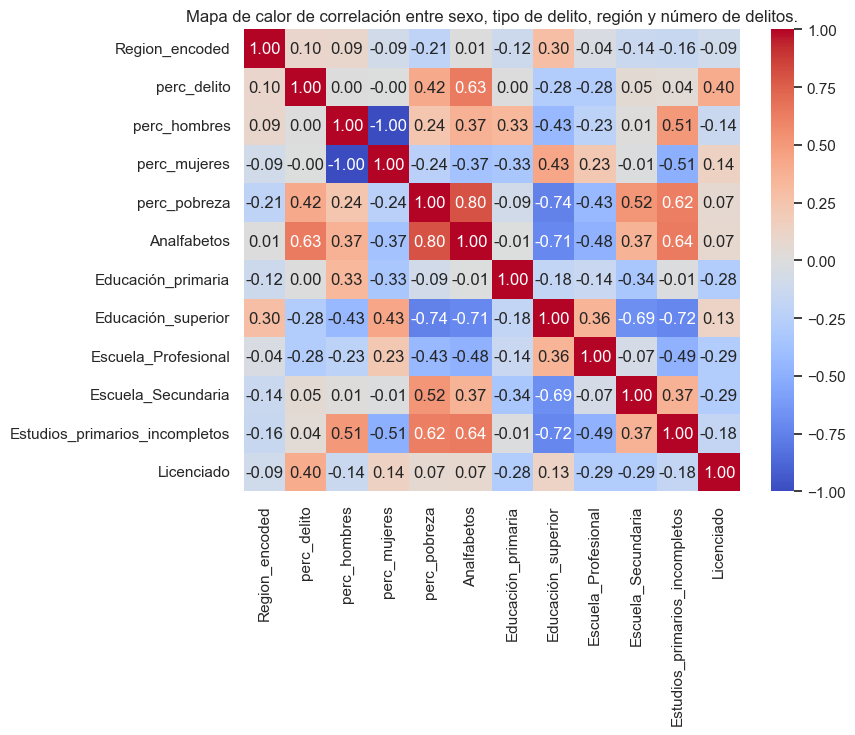
The below scatter plot gives the impression that Educación primaria and Escuela\_Profesional has very less correlation and so has to be most likely ecluded from the modelling.

s

1. Heat Map to see correlation among all the variables

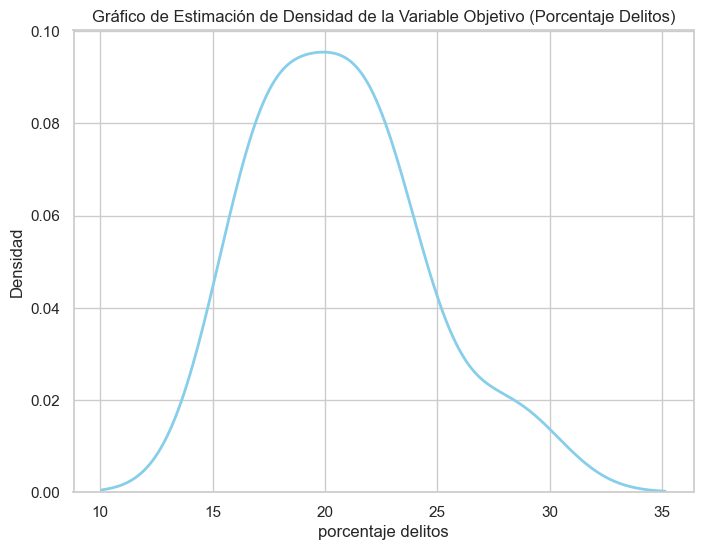
Based on the heat map following conclusions were arrived

* porcentaje Delitos has strong positive and negative relationship based on education level.
* porcentaje Delitos and poverty porcentaje lso has a good positive correlaltion
* porcentaje Delitos and genero also has little correlation and so should be dropped
* Few categories with in education classes got less correlation and so will be dropped as this can negatively impact model outcome. The categories are Educación primaria and Escuela\_Profesional.



1. Density plot for crime percentage

The Density Estimation Graph of the Target Variable (Percentage of Crimes) shows almost a normal distribution which makes the field ideal for prediction.



1. Plotting Percentage of crime against year

It's evident the crime rate during 2020 was less due to corona from the below picture.



**Explanation of models, Measures to adapt model and Visualization**

1. Modelling for total Crime percentage by year

Here we will first try to predict the total crime percentage by year . Region parameter is excluded from the modeling here. This prediction and plotting is to get an idea on the overall pattern of crime across Spain.

In this instance, the ARIMA model is employed for prediction purposes, with the parameters p, q, and d determined through the use of auto\_arima, resulting in the configuration (1,2,0). With an ADF Statistic of -740.325475 and a p-value of approximately 0.000000, you have strong statistical evidence to reject the null hypothesis of the ADF test. This means it's statistically significant that the time series does not have a unit root, implying that the series is stationary which makes it ideal for forecasting.

A graph with blue lines and red dots

Description automatically generated

The forecasted result for three future timesteps are

A black background with white text

Description automatically generated

Following are the model summary statistics.

A screenshot of a computer screen

Description automatically generated

* 1. Forecasting crime percentage across region using Prophet model

Prophet uses a decomposable time series model with three main components: trend, seasonality, and holidays. For the trend, it typically uses a piecewise linear or logistic growth curve to model non-periodic changes in the value of the time series. We have created separate models for each region and use this in prediction

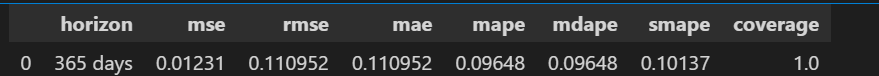
Below picture shows Plotting the forecast components for one of the regions. The spread shows the the prediction range is very less and implies the accuracy of model. We have used both additive and multiplicative regressors in the model.



The trend line is increasing over time, indicating a prediction of upward movement in the metric being forecasted.

The additive extra regressors suggest that the additional factors considered in the model that are having a constant additive effect on the forecast. This effect does not seem to change over the time period displayed.

There is categorical encoding specifically done for region field before the data to the model . Prophet model model came with a error metrics showing root means squared error as 0.110952 which shows that the model is adapted well for this dataset.

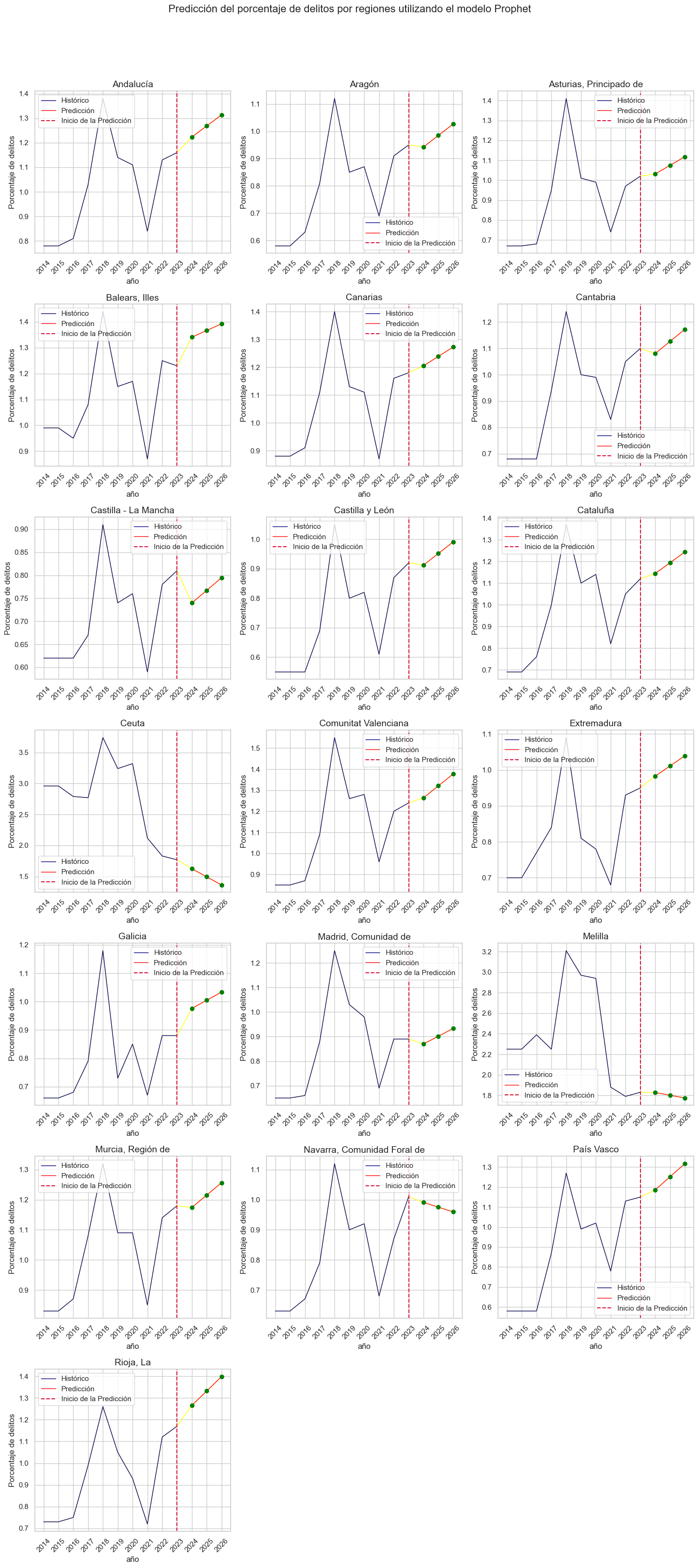


Below diagram shows the mean coefficients regressor values across all regions.So for every unit increase in "Analfabetos," the forecasted target is expected to increase by about 1.617137 units

A screenshot of a computer

Description automatically generated

Below plot shows the predictions done for each region by Prophet model

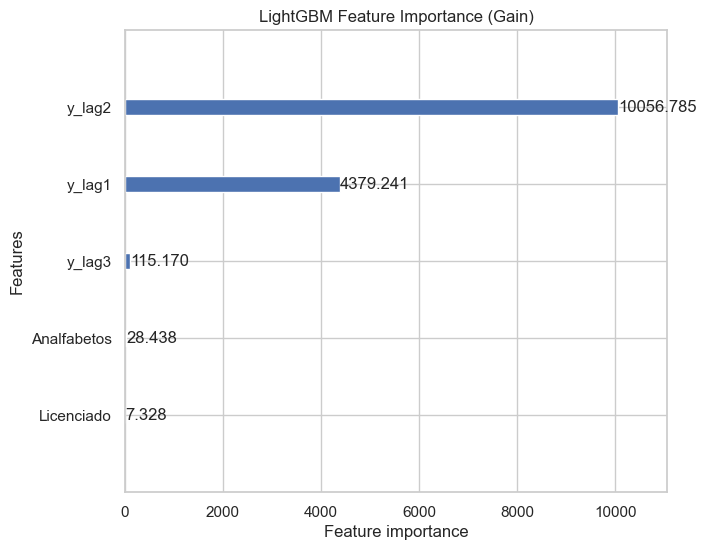


1. Forecasting crime percentage by year and other regressors using Light GBM regressor model

LightGBM regressor is a gradient-boosting framework that uses tree-based learning algorithms. For time series forecasting with LGBM, we transformed time series data into a supervised learning problem. This involves creating lag features (past values) as predictors.Given the absence of time series elements in this context, we introduced three lagged features from previous time steps for each record category. This approach ensures the establishment of a connection with past time steps. Additionally, label encoders were used to transform the data prior to its application to the model. The learning rate was adjusted to explore different results, ultimately being fixed at 0.001. The resulting Root Mean Square Error (RMSE) was approximately 0.1374.

Predictions were made across all combinations of region for three future steps.

The below picture shows the importance of each feature on the model prediction. The exogeneous regressors has very importance in overall prediction.



The chart below presents the prediction results with the crime percentage.

A screenshot of a graph

Description automatically generated

1. Forecasting crime count by year and other regressors using Linear regression model

Linear regression is a common machine learning technique used to predict a dependent variable based on one or more independent variables

Due to the lack of time series elements in the dataset, we incorporated three previous time steps as lagged variables for each record category, enabling the establishment of connections based on historical data. Label encoders were also used to preprocess the data before fitting it into the model. The model achieved a Root Mean Square Error (RMSE) of approximately 0.2604. Predictions were executed across various combinations of sex extending two steps into the future.

Below picture shows the feature coefficients for linear regression.

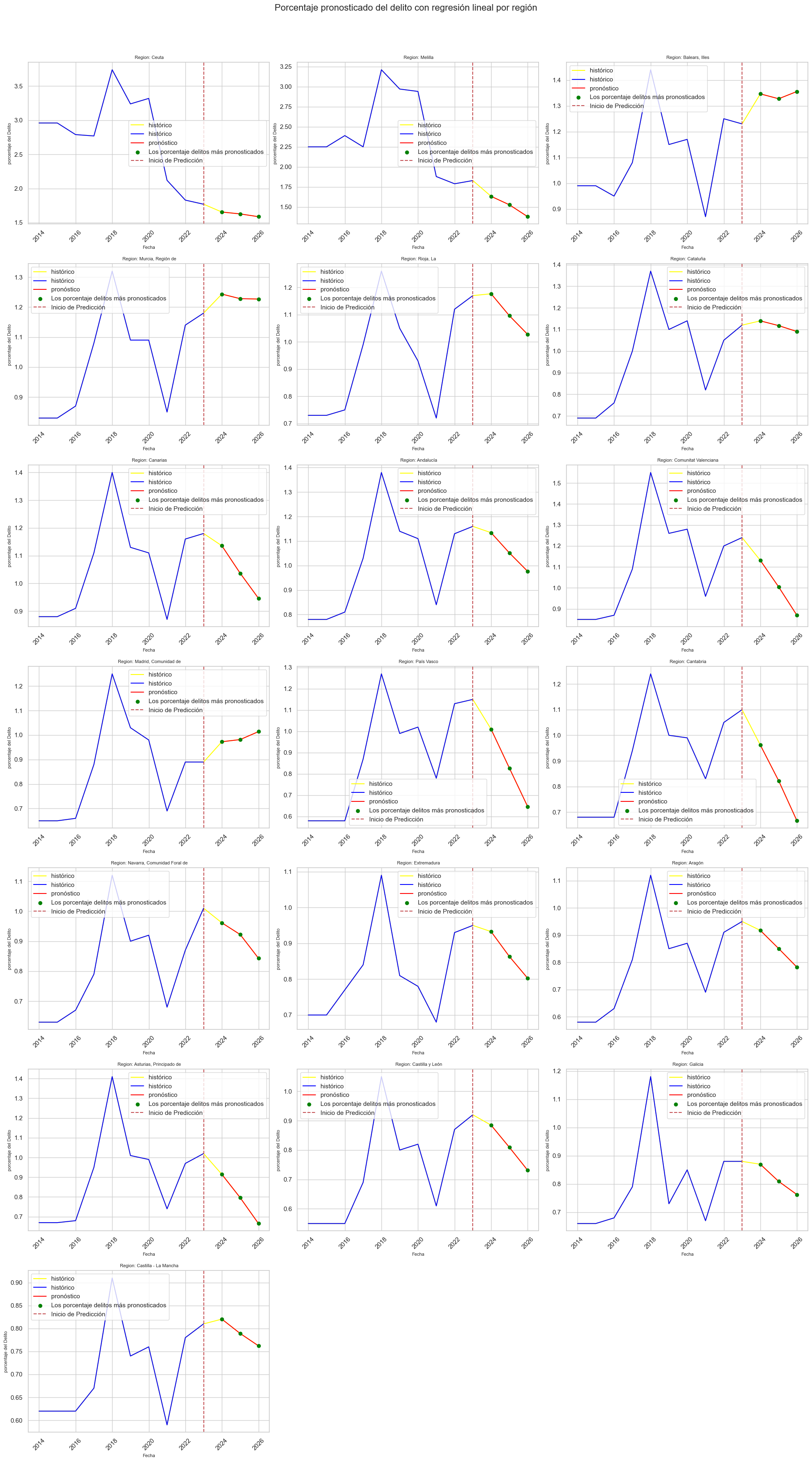
A graph with blue bars

Description automatically generated

Upward bars (above 0 on the y-axis) denote a positive relationship, where increases in the feature value are associated with increases in the predicted target value.

Downward bars (below 0 on the y-axis) denote a negative relationship, where increases in the feature value are associated with decreases in the predicted target value.

The chart provided below displays prediction outcomes for all the regions.



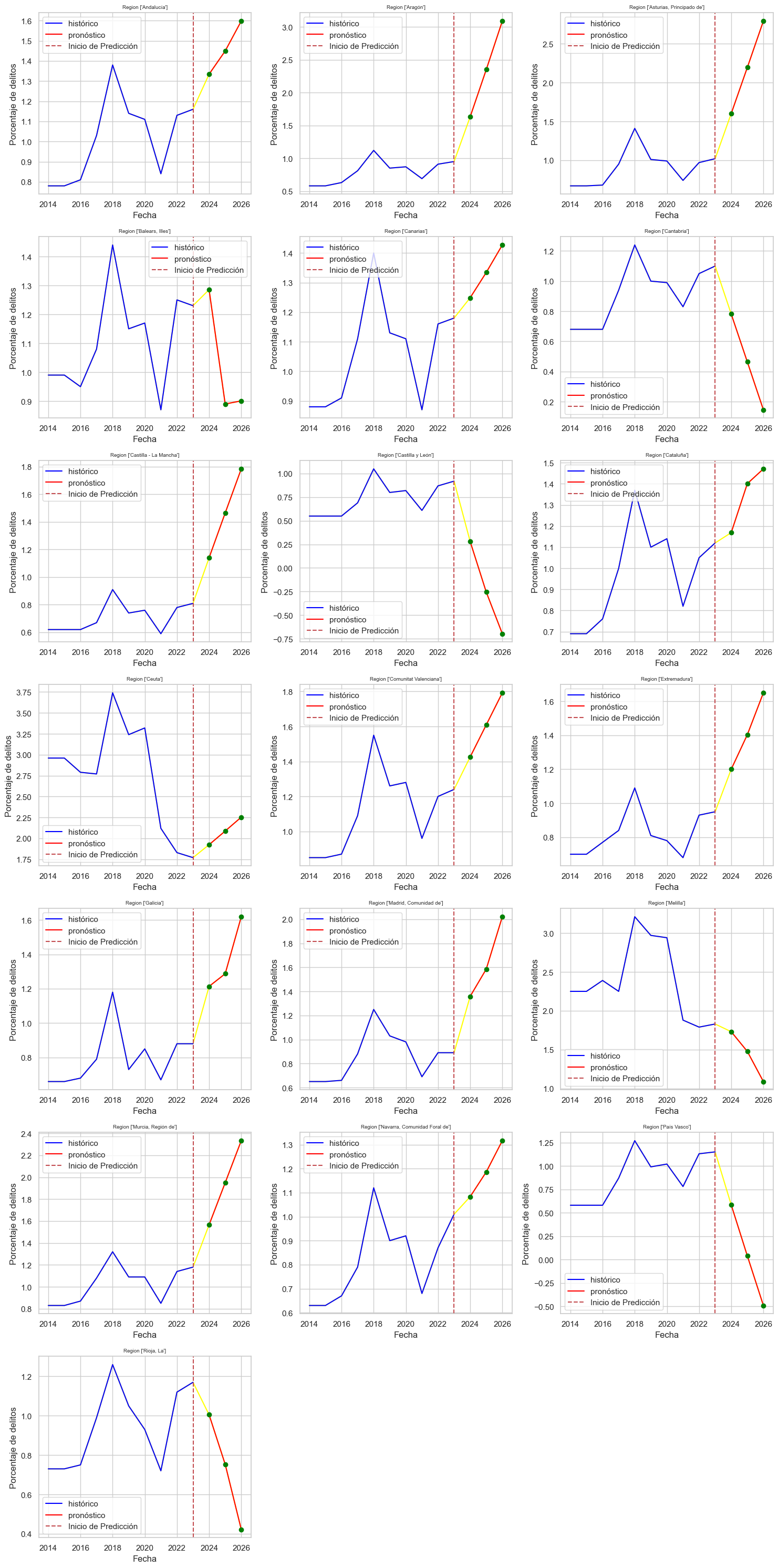
1. Forecasting crime count by year and other regressors applying ARIMA across individual Region categories

ARIMA, which stands for AutoRegressive Integrated Moving Average, is a statistical analysis model used for time series forecasting.

Here we took a distinct approach of training only the time components specific to each region to create a model and thereby created 19 models in total( one for each region). This is to also see how the individual models perform in comparison to combined training which was done before.

To evaluate the model's predictive performance, a preliminary analysis was conducted using ARIMA on each category within a single region. The ARIMA model was specified with an order of (1,2,0) for (p,d,q).

Even though the model was added with external regressors( education and poverty), a maximum likelihood convergence issue occurred. This means that the optimization algorithm used to estimate the parameters of the ARIMA model is unable to find the best set of parameters given the maximum likelihood criterion. The most likely case here for us is the complex seasonality. The maximum iterations parameters are set to 1000 to overcome this.



Below is the model summary for one of the regions( Rioja La). From the P value(P>|Z|), it appears like Analfabetos and Escuela\_Profesional coefficients have p-values of 0.000 and 0.002, respectively, indicating they are statistically significant as it is above .05 thereshold. Note this cannot be generalised as we have created separate models for each region

**A screenshot of a computer screen

Description automatically generated**

**Final Conclusion:**

Among the models tested, the Prophet model, which falls under the decision tree category, emerged as the top performer, delivering the best results. Also Light GBM regression and Linear regression also faired well with favorable outcomes. However the effect of external regressions variables had little effect on prediction for those.The additional features variables selected in our model creation phase impacts the overall accuracy and effectiveness of the results to a certain degree.

**Further Possibilities**

* Employing Ridge or Lasso regression could enhance the predictive performance and accuracy of the linear regression model.
* Adopting probabilistic modeling with PYMC could substantially expand the dataset. Introducing stochastic variables and incorporating monthly increments may facilitate this expansion.
* Leveraging deep learning approaches, such as LSTM or Transformer models, has the potential to yield improved results compared to current models.