# Proyecto de Minería de Datos

## Portada


- **Nombre del Analista:** Alejandro Borrego Megías
- **Fecha:** 24-12-2023
- **Correo Electrónico:** alejbormeg@gmail.com


## Índice



## 1. Introducción al objetivo del problema y las variables implicadas.

En el marco de la investigación y análisis demográfico y político, se aborda el desafío de comprender y prever los patrones de abstención en las elecciones municipales en España. La abstención electoral, medida a través del porcentaje

de abstención, es una variable crucial que refleja la participación ciudadana en el proceso democrático.

El conjunto de datos utilizado, denominado "DatosEleccionesEspaña.xlsx", contiene información demográfica detallada sobre los municipios de España, así como los resultados de las últimas elecciones. Este conjunto incluye variables que abarcan desde características poblacionales hasta resultados de votación, proporcionando una visión integral de los factores que podrían influir en la abstención.

El objetivo principal de este análisis es desarrollar dos modelos predictivos: uno de regresión lineal para predecir el porcentaje de abstención y otro de regresión logística para prever la probabilidad de una alta abstención. Estos modelos tienen el propósito de identificar patrones y relaciones significativas entre las diversas variables demográficas y los resultados electorales, lo que podría ayudar a comprender mejor los factores que afectan la participación electoral.

Las variables consideradas serán las siguientes:

| Variable | Descripción |
|---|---|
| Name | Nombre del municipio |
| CodigoProvincia | Código de la provincia (coincide con los dos primeros dígitos del código postal). Toma 52 valores distintos |
| CCAA | Comunidad autónoma a la que pertenece el municipio |
| Population | Población del municipio en 2016 |
| TotalCensus | Población en edad de votar en 2016 |
| AbstencionAlta | Variable dicotómica que toma el valor 1 si el porcentaje de abstención es superior al 30%, y 0 en otro caso |
| AbstentionPtge | Porcentaje de abstención |
| Age_0-4_Ptge | Porcentaje de ciudadanos con menos de 5 años |
| Age_under19_Ptge | Porcentaje de ciudadanos con menos de 19 años |
| Age_19_65_pct | Porcentaje de ciudadanos entre 19 y 65 años |
| Age_over65_pct | Porcentaje de ciudadanos con más de 65 años |
| WomanPopulationPtge | Porcentaje de mujeres |
| ForeignersPtge | Porcentaje de extranjeros |
| SameComAutonPtge | Porcentaje de ciudadanos que reside en la misma provincia en la que nacieron |
| SameComAutonDiffProvPtge | Porcentaje de ciudadanos que reside en la misma CCAA en la que nacieron, pero distinta provincia |
| DifComAutonPtge | Porcentaje de ciudadanos que reside en la distinta CCAA de la que nacieron |
| UnemployLess25_Ptge | Porcentaje de parados de menos de 25 años |
| Unemploy25_40_Ptge | Porcentaje de parados entre 25 y 40 años |
| UnemployMore40_Ptge | Porcentaje de parados de más de 40 años |
| AgricultureUnemployPtge | Porcentaje de parados en el sector de la agricultura |
| IndustryUnemployment Ptge | Porcentaje de parados en el sector de la industria |

| Variable | Descripción |
|---|---|
| ConstructionUnemployment_Pct | Porcentaje de parados en el sector de la construcción |
| ServicesUnemployment_Pct | Porcentaje de parados en el sector servicios |
| totalEmpresas | Número total de empresas en el municipio |
| Industria | Número de empresas del sector industrial en el municipio |
| Construccion | Número de empresas del sector de la construcción en el municipio |
| ComercTTEHosteleria | Número de empresas dedicadas a comercio, transporte u hostelería en el municipio |
| Servicios | Número de empresas del sector servicios en el municipio |
| ActividadPpal | Actividad principal de las actividades del municipio (Industria, Construcción, ComercTTEHosteleria, Servicios y Otros) |
| inmuebles | Número de inmuebles en el municipio |
| Pob2010 | Población en el municipio en 2010 |
| SUPERFICIE | Superficie del municipio |
| densidad | Densidad de población del municipio: MuyBaja (<1 hab/ha), Baja (entre 1 y 5 hab/ha), Alta (>5 hab/ha) |
| PobChange_pct | Porcentaje de cambio en la población (valores negativos indican que ha disminuido). Respecto a las anteriores elecciones |
| PersonasInmueble | Número medio de personas que habita un inmueble |
| Explotaciones | Número de explotaciones agrícolas en el municipio |

En última instancia, este estudio busca proporcionar información valiosa para entender los determinantes de la participación electoral en los municipios españoles, contribuyendo así a la toma de decisiones informada en el ámbito político y social.

## 2. Importación del conjunto de datos y asignación correcta de los tipos de variables.

La base de datos se guarda en la carpeta `src/data` y se realiza la importación del conjunto de datos con la librería `Pandas` de Python. Una vez hecho esto eliminamos las variables objetivo relacionadas con el porcentaje de Izquierda, Derecha y otros, tanto continuas como categóricas:

```
# Cargo los datos
datos = pd.read_excel('src/data/DatosEleccionesEspana.xlsx')

# Eliminamos las variables que no usaremos
variables_a_eliminar = ["Izda_Pct", "Dcha_Pct", "Otros_Pct", "Izquierda", "Derecha"]

datos = datos.drop(columns=variables_a_eliminar)
```

Comprobamos que todas las variables tienen los tipos correctos ejecutando:

```
print(datos.dtypes)
```

Obteniendo:

```
Name                              object
CodigoProvincia                    int64
CCAA                              object
Population                         int64
TotalCensus                        int64
AbstentionPtge                   float64
AbstencionAlta                     int64
Age_0-4_Ptge                     float64
Age_under19_Ptge                 float64
Age_19_65_pct                    float64
Age_over65_pct                   float64
WomanPopulationPtge              float64
ForeignersPtge                   float64
SameComAutonPtge                 float64
SameComAutonDiffProvPtge         float64
DifComAutonPtge                  float64
UnemployLess25_Ptge              float64
Unemploy25_40_Ptge               float64
UnemployMore40_Ptge              float64
AgricultureUnemploymentPtge      float64
IndustryUnemploymentPtge         float64
ConstructionUnemploymentPtge     float64
ServicesUnemploymentPtge         float64
totalEmpresas                    float64
Industria                        float64
Construccion                     float64
ComercTTEHosteleria              float64
Servicios                        float64
ActividadPpal                     object
inmuebles                        float64
Pob2010                          float64
SUPERFICIE                       float64
Densidad                          object
PobChange_pct                    float64
PersonasInmueble                 float64
Explotaciones                      int64
```

Como vemos, las variables categóricas (Name, CCAA, ActividadPpal, Densidad) tienen el tipo `object` correctamente, mientras que las demás son numéricas todas, algunas enteros y otras en coma flotante.

## 3. Análisis descriptivo del conjunto de datos.

Ejecutando `datos.shape` observamos que las dimensiones del dataframe cargado son de `(8119, 36)`, lo que implica un total de 8119 ejemplos en la base de datos y un total de 36 variables (incluyendo las variables objetivo) que analizar y limpiar.

Separamos las variables en variables numéricas y categóricas:

```
# Seleccionar las columnas numéricas del DataFrame
numericas = datos.select_dtypes(include=['int', 'int32', 'int64','float', 'float32', 'float6

# Seleccionar las columnas categóricas del DataFrame
categoricas = [variable for variable in variables if variable not in numericas]
```

Obtenemos un total de 32 variables numéricas y 4 categóricas. Tras esto procedemos a un análisis más exhaustivo de las distintas variables. Para las variables categóricas emplearemos la función `analizar_variables_categoricas` del fichero `src/FuncionesMineria.py`. Esta función nos devuelve para cada variable categórica el número de ocurrencias para cada categoría así como el porcentaje que representa dentro del total de datos:

```
# Frecuencias de los valores en las variables categóricas
analisis_categoricas = analizar_variables_categoricas(datos)

print(analisis_categoricas)
```

El resultado obtenido es el siguiente:

- Número y naturaleza de variables
- Datos erróneos, etc.

## 4. Corrección de los errores detectados.

## 5. Análisis de valores atípicos.

- Decisiones tomadas.

## 6. Análisis de valores perdidos.

- Estrategias de imputación.

## 7. Transformaciones de variables y relaciones con las variables objetivo.

## 8. Detección de las relaciones entre las variables input y objetivo.

## 9. Construcción del modelo de regresión lineal.

- Selección de variables clásica
- Selección de variables aleatoria
- Selección del modelo ganador
- Interpretación de los coeficientes de dos variables incluidas en el modelo (una binaria y otra continua)
- Justificación del mejor modelo y medición de la calidad del mismo

## 10. Construcción del modelo de regresión logística.

```
- Selección de variables clásica
- Selección de variables aleatoria
- Selección del modelo ganador
- Determinación del punto de corte óptimo
- Interpretación de los coeficientes de dos variables incluidas en el modelo (una binaria y
- Justificación del mejor modelo y medición de la calidad del mismo
```