**TALLER CIENCIA DE DATOS**

**Datasets a utilizar: -**Las bases de datos que utilizaremos son las siguientes: los resultados de las últimas elecciones nacionales del 2014, los cuales son públicos y se encuentran en la siguiente página web: <https://catalogodatos.gub.uy/dataset/corte-electoral-resultado-elecciones-nacionales-2014>

Y la encuesta continua de hogares que realiza el Instituto Nacional de Estadística (INE) de la que se dispone públicamente en su sitio web: <https://catalogodatos.gub.uy/dataset/ine-encuesta-continua-de-hogares-ano-2014>

**Problemáticas a analizar:**

Suponiendo poca movilidad respecto a los votos registrados en las elecciones anteriores y analizando los datos por circuito se pueden captar determinadas características de preferencia según la disposición geográfica a nivel de barrios, logrando aproximarse a la definición de perfil de votante.

De esta forma se logra una variable objetivo que es el porcentaje que obtuvo cada partido o candidato, en el marco de una serie de atributos obtenidos de la encuesta continua de hogares.

Así puede completarse el perfil de la población en determinado territorio (ingresos promedio de ese circuito, porcentaje de ricos, porcentaje de pobres, porcentaje de mujeres, de hombres, de personas casadas, etc.) y se relaciona con los resultados electorales.

**Objetivos:**

Inferir la probabilidad de éxito electoral por circuito para definir las estrategias de campaña política para elegir sobre qué grupos actuar: dónde están los indecisos, dónde no vale la pena desplegar acciones de campaña.

Por ejemplo, si un candidato obtuvo mayor porcentaje de hombres que de mujeres y hay un pueblo del interior en el que hay muchos más hombres que mujeres, podría inferirse que hay mayor probabilidad de éxito desplegando acciones de campaña en dicho pueblo.

Es recomendable tener un ambiente de training y otro de test donde el modelo se realice en base al dataset de training y pueda ser probado en test para calcular el porcentaje de prediccion del modelo realizado

Tener elecciones de años anteriores ejemplo 2004 2009 para poder predecir el 2014 y luego tener los tres años y predecir el 2019.

Antes de empeza a realizar predicciones analisar los datos y buscar relaciones entre ellos y datos mas interesantes

Datos disponibles, Elecciones nacionales 2014

Encuesta Nacional de hogares 2014 ( mas reciente 2017)

Predecir votacion por departamento de elecciones 2004-2009-2014 y la proxima de 2019

**pasos a seguir**

-terminar de mapear circuitos con barrios

-Definir archivos del dataset

-Separar datos para crear dataset de training y test

-Analizar relaciones entre atributos y datos interesantes

-Ver posibles predicciones por departamento y por barrio para elecciones 2019

**Datos interesantes por documento:**

Hogares(2014):

* Estrato por Barrio
  + destacar 3-5 barrios de mayor porcentaje por estrato
* CANTIDAD DE HOMBRES
* CANTIDAD DE MUJERES
* CANTIDAD DE MENORES DE 14 AÑOS
* CANTIDAD DE MAYORES DE 60 AÑOS
* CANTIDAD DE OCUPADOS
* CANTIDAD DE DESOCUPADOS
* Promedio ingreso hogar por barrio

Personas 2014 (se toman en cuenta solamente personas mayores de 18 años):

* Estrato por Barrio
  + destacar 3-5 barrios de mayor porcentaje por estrato
* Franjas Etarias por barrio
* Porcentaje de ascendencia(principal) por barrio
* porcentaje estado civil por barrio.
* Nivel educativo por barrio.
* condición de actividad por barrio
* Promedio de ingresos (por persona) por barrio

Elecciones votos 2014:

* todos los porcentajes de votos por departamento

Elecciones por partido 2014:

* cantidad de votos por partido ( por departamento)

Cruce de datos:

* Porcentaje de votos por barrio
* Porcentaje por partido por barrio
* Promedio ingreso hogar por partido
* Estrato de persona por barrio y partido (correlación)

Luego de realizar los porcentajes de los datos de las personas analizar correlación con posible partido

Para modelo predictivo utilizaremos las votaciones de 2004-2009-2014 por departamento y por partido e intentaremos predecir elecciones 2019

Nota: cuando se habla de barrios solo se tomarán en cuenta valores de montevideo.