Aplicación de modelos de regresión y técnicas de clustering en datos de COVID -19

Ambrogi, Tomás[1]; Rocamora, Leandro[1] ; Suli, Solange [2]

[1] Universidad Tecnologica Nacional, Facultad Regional Buenos Aires
[2] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales

Introducción

- La enfermedad por coronavirus 2019 (COVID-19) es una enfermedad infecciosa causada por un síndrome respiratorio severo coronavirus 2 (SARS-CoV-2).
- Según el Reporte Diario Vespertino Nro 389, del **30 de Septiembre del 2020** emitido por Ministerio de Salud de la República Argentina fueron registrados, **para la Ciudad Autónoma de Buenos Aires (CABA), 125.068 casos positivos hasta el día de la fecha**.
- **Objetivos:**

  → Entrenar, escoger y evaluar un modelo de regresión que permita **predecir cantidad de fallecidos del día siguiente por coronavirus.**

  → Estudiar la relación de los datos a través de aprendizaje no supervisado usando algoritmos de clustering para evaluar la existencia de segmentación de los datos.

Datasets

- Base de datos del Ministerio de Salud de la República Argentina **(mar-sept 2020)** → filtrado de casos registrados para **CABA** y se tomó como fecha de referencia del caso la "fecha de apertura" ⇒ dataset: 18 features y 361133 samples.

Análisis Exploratorio de Datos

- Tratamiento de datos faltantes:
  - features "fechas.." → se transforman en variables booleanas (ej: "fecha_inicio_sintomas" ⇒ "presenta_sintomas").
  - "edad" (37 datos faltantes) → datos descartados.
- Creación de nuevas features de interés ("argentino", "mayor_65", "publico", entre otras).
- Acumulación de casos diarios (según "fecha_apertura"), confirmados de CABA a partir de una Tabla Pivot.

Métodos

- **Clustering:**
  - Se trabajó con el **dataset original** así como también con sus **componentes principales**.
  - el dataset original se descompone en otros dos datasets:
    i. 30 componentes principales
    ii. 6 componentes principales.

Cada uno de los datasets es sometido al mismo proceso de clustering por medio del algoritmo **KMeans.** Se estudian los resultados en base a 2, 4 y 6 clusters.

- **Regresión:**
  - Modelos de regresión planteados: Linear Regression, **Ridge Regression**, Support Vector Machines y K Nearest Neighbors Regressor.
  - Entrenamiento: **datos por día desde 01/03/2020 hasta 30/09/2020**
  - Validación: primeros quince días de septiembre.
  - Escalado normal de los datos → μ =0 y σ=1.
  - Media móvil (ventana de 2 días) ⇒ suavizar y reducir la variabilidad de los datos de entrada.
  - **Evaluación y Métricas**
    - Para garantizar la elección del mejor modelo → validación cruzada → **TimeSeriesSplit** (por temporalidad de los datos).
    - **Maximizar** el **R2**, y al mismo tiempo, **minimizar** tanto el **MAE** como el **RMSE**.

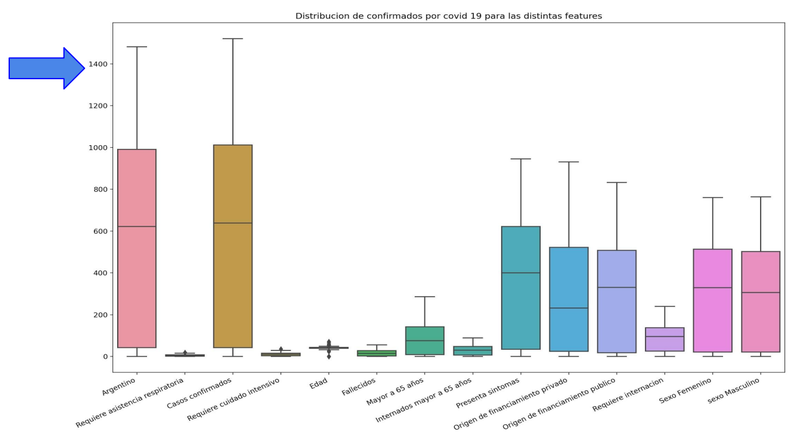

Figura 1: Distribución de casos confirmados para las distintas features obtenidas.

Resultados

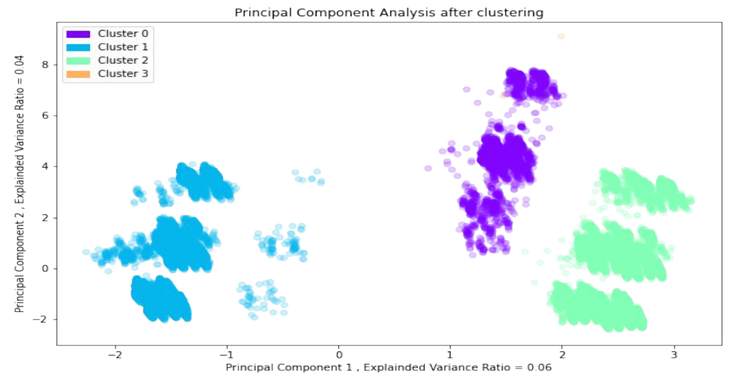

Figura 2: Clusters graficados sobre las dos componentes principales

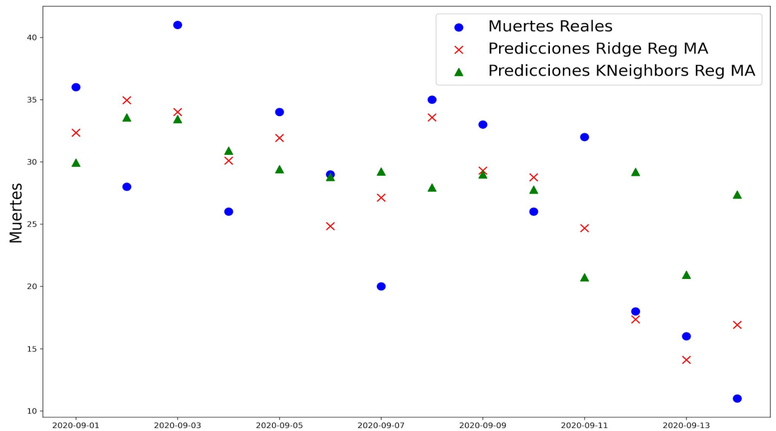

Figura 3: gráfico de dispersión en el que se detallan: las muertes reales (en azul), las predicciones con el mejor modelo (en rojo) y las predicciones con el peor modelo (en verde) para las fechas del 1 al 14 de septiembre.

Conclusiones

- En clustering se observó que definir 4 o más clusters genera al menos un cluster vacío, indicando que para segmentar los datos **3 o menos clusters son suficientes**.
- Se encontró una **relación entre las variables de entrada y los fallecidos del día siguiente** que se clarifica al realizar una **media móvil uniforme de dos días**.