# Enfermedad Dengue en la República Argentina

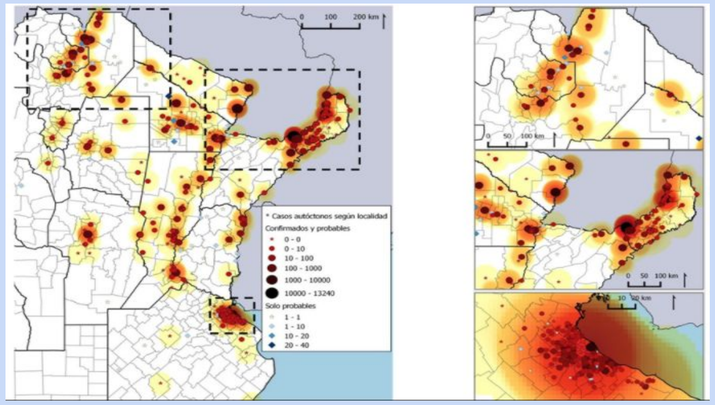
## Aplicación de Machine Learning para su estudio y predicción de casos

**Fresca Lorenzo [1], Mojico Ailín[1] , Rosselló Matías[1]**

*[1]Universidad Tecnológica Nacional, Facultad Regional Buenos Aires*

## Introducción

El virus del dengue afecta a las zonas con climas tropicales y subtropicales. Para el estudio y predicción de casos de dengue en la República Argentina se utilizaron los registros de casos de los últimos 3 años y los registros de cada centro de medición del Servicio Meteorológico Nacional.




El objetivo es poder predecir la cantidad de casos de dengue que puedan contabilizarse en cada provincia argentina, según determinadas variables meteorológicas y basándonos en los casos históricos.

## Datasets

Los datasets se obtuvieron del portal de datos de la Nación. Además, se solicitaron nuevos datos al Ministerio de Salud y al Servicio Meteorológico Nacional.

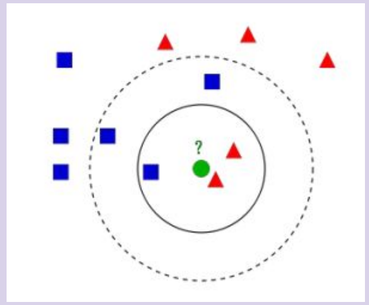
| Dataset | Significado samples | Año | | |
|---|---|---|---|---|
| | | 2018 | 2019 | 2020 |
| *Vigilancia_dengue* | Reportes semanales de casos de Dengue | **679** | **807** | **9.002** |
| *Exp_horarios* | Mediciones de meteorológicas cada 6 horas por estación | **97.690** | **98.179** | **48.128** |
| *Exp_precipitaciones* | Mediciones de precipitaciones mensuales por estación | **840** | **840** | **840** |
| Exp_observatorios | Estaciones y su localización | **70** | | |

## Métodos

Se realizaron 7 métodos distintos de Machine Learning: Regresión lineal, Ridge Regression (L1), Lasso (L2), Super Vector Regression (SVR), KNN Regression, Regresión Polinómica y Random Forest Regression.

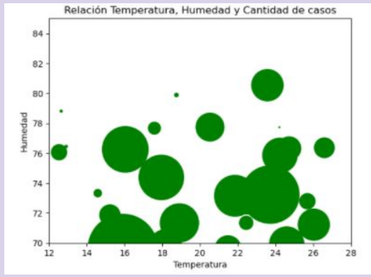
Particularmente para KNN Regression, para definir el valor de K se aplicó un proceso de iteración. Desde K=0 hasta K=15. El valor que mejor resultado trajo fue K=2.

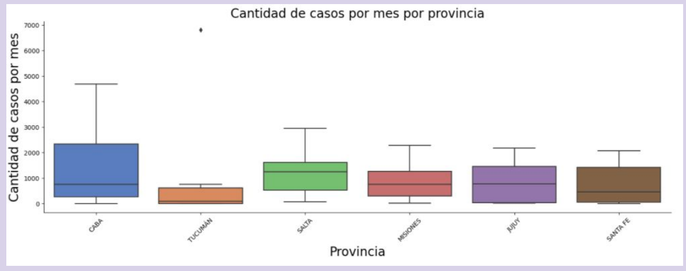
Con el parámetro K definido, se entrenó el modelo y luego se testeó calculando, a raíz de esto el MSE, el MAE y el R2 .




## Análisis Exploratorio de Datos

Se tomaron las features de temperatura, humedad y cantidad de casos y se observó la relación entre ellas a través de un gráfico de dispersión. Cuanto más grande el círculo, mayor la cantidad de casos.




## Resultados

El modelo KNN con K=2, resultó ser el algoritmo de aprendizaje más preciso y con menores errores de predicción.

| Modelo | R2 | RMSE | MAE |
|---|---|---|---|
| KNN (K=2) | 0.320381 | 99391.862500 | 95.441667 |
| SVR | 0.317942 | 99748.575227 | 164.337727 |
| Random Forest Regressor | 0.241314 | 110955.225157 | 124.674517 |
| Lasso | 0.062751 | 137069.401930 | 231.725174 |
| Ridge Regression | 0.060763 | 137360.102330 | 232.822357 |
| Linear | 0.056678 | 137957.510123 | 234.022931 |
| Polynomial Regression | -3.246194 | 788.029767 | 620990.913268 |

## Conclusiones

No se encontró una relación directa entre las variables meteorológicas y la cantidad de casos. Asimismo, no se descarta que exista una relación adicionando variables como la flora y fauna de cada región.

Se observa que en los primeros seis meses del año son los que engloban más del 98% de los casos reportados en un año, con un pico entre los meses de marzo y mayo.