1. ¿Puedes predecir las epidemias locales de dengue?

La fiebre del dengue es una enfermedad transmitida por mosquitos que ocurre en partes tropicales y subtropicales del mundo. En casos leves, los síntomas son similares a los de la gripe: fiebre, erupción cutánea y dolor muscular y articular. En casos severos, la fiebre del dengue puede causar sangrado severo, presión arterial baja e incluso la muerte.  
  
Debido a que es transmitido por mosquitos, la dinámica de transmisión del dengue está relacionada con variables climáticas como la temperatura y la precipitación. Aunque la relación con el clima es compleja, un número creciente de científicos argumenta que es probable que el cambio climático produzca cambios distributivos que tendrán importantes implicaciones para la salud pública en todo el mundo.  
  
En los últimos años, la fiebre del dengue se ha extendido. Históricamente, la enfermedad ha sido más frecuente en el sudeste asiático y las islas del Pacífico.

Utilizando los datos ambientales recopilados por varias agencias del Gobierno Federal de los EE. UU., Desde los Centros para el Control y la Prevención de Enfermedades hasta la Administración Nacional Oceánica y Atmosférica del Departamento de Comercio de los EE. UU, ¿Puede predecir el número de casos de fiebre del dengue reportados cada semana en San Juan, Puerto Rico e Iquitos, Perú?

Esta es una competencia de práctica de nivel intermedio. Su tarea es predecir el número de casos de dengue cada semana (en cada ubicación) en función de las variables ambientales que describen los cambios de temperatura, precipitación, vegetación y más.

La comprensión de la relación entre el clima y la dinámica del dengue puede mejorar las iniciativas de investigación y la asignación de recursos para ayudar a combatir las pandemias que amenazan la vida.

# Predecir la próxima iniciativa pandémica Los datos para esta competición provienen de múltiples fuentes destinadas a apoyar la iniciativa “Predict the Next Pandemic”. Los datos de vigilancia del dengue son proporcionados por los Centros para el Control y la Prevención de Enfermedades de EE. UU., así como por la Unidad 6 de Investigación Médica Naval del Departamento de Defensa y el Centro de Vigilancia Sanitaria de las Fuerzas Armadas, en colaboración con el gobierno peruano y las universidades de EE. UU. Los datos ambientales y climáticos son proporcionados por la Administración Nacional Oceánica y Atmosférica (NOAA), una agencia del Departamento de Comercio de los Estados Unidos. En sus propias palabras: *Las predicciones precisas del dengue ayudarían a los trabajadores de salud pública ... y a las personas de todo el mundo a tomar medidas para reducir el impacto de estas epidemias. Pero predecir el dengue es una tarea considerable que requiere la consolidación de diferentes conjuntos de datos sobre la incidencia de enfermedades, el clima y el medio ambiente.* Este es un problema complicado, sin duda. Pero los datos reales a menudo son complicados y desordenados. Estudie utilizando los siguientes recursos, ¡sus ideas podrían salvar vidas!

* [Dengue Forecasting Homepage](http://dengueforecasting.noaa.gov/)
* [CDC Dengue Overview](http://www.cdc.gov/Dengue/)
* [NOAA Wiki](https://en.wikipedia.org/wiki/National_Oceanic_and_Atmospheric_Administration)

# Descripción del problema

# Su objetivo es predecir el valor de la columna *total\_cases* para cada (ciudad, año, semana del año) en el conjunto de prueba. Hay dos ciudades, San Juan e Iquitos, con datos de prueba para cada ciudad que abarcan 5 y 3 años respectivamente. Realizará una entrega que contenga predicciones para ambas ciudades. Los datos para cada ciudad se han concatenado junto con una columna de ciudad que indica la fuente (por ejemplo sj para San Juan e iq para Iquitos). El conjunto de prueba es un futuro puro, lo que significa que los datos de la prueba son secuenciales y no se superponen con ninguno de los datos de entrenamiento. En todo momento, los valores faltantes se han rellenado como NaN.

## The features in this dataset

You are provided the following set of information on a (year, weekofyear) timescale:

(Where appropriate, units are provided as a \_unit suffix on the feature name.)

### City and date indicators

* city – City abbreviations: sj for San Juan and iq for Iquitos
* week\_start\_date – Date given in yyyy-mm-dd format

### NOAA's GHCN [daily climate data](https://www.ncdc.noaa.gov/oa/climate/ghcn-daily/) weather station measurements

* station\_max\_temp\_c – Maximum temperature
* station\_min\_temp\_c – Minimum temperature
* station\_avg\_temp\_c – Average temperature
* station\_precip\_mm – Total precipitation
* station\_diur\_temp\_rng\_c – Diurnal temperature range

### PERSIANN [satellite precipitation measurements](http://www.ncdc.noaa.gov/cdr/operationalcdrs.html) (0.25x0.25 degree scale)

* precipitation\_amt\_mm – Total precipitation

### NOAA's NCEP [Climate Forecast System Reanalysis](http://rda.ucar.edu/datasets/ds093.0/#metadata/detailed.html?_do=y) measurements (0.5x0.5 degree scale)

* reanalysis\_sat\_precip\_amt\_mm – Total precipitation
* reanalysis\_dew\_point\_temp\_k – Mean dew point temperature
* reanalysis\_air\_temp\_k – Mean air temperature
* reanalysis\_relative\_humidity\_percent – Mean relative humidity
* reanalysis\_specific\_humidity\_g\_per\_kg – Mean specific humidity
* reanalysis\_precip\_amt\_kg\_per\_m2 – Total precipitation
* reanalysis\_max\_air\_temp\_k – Maximum air temperature
* reanalysis\_min\_air\_temp\_k – Minimum air temperature
* reanalysis\_avg\_temp\_k – Average air temperature
* reanalysis\_tdtr\_k – Diurnal temperature range

### Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's [CDR Normalized Difference Vegetation Index](https://www.ncdc.noaa.gov/cdr) (0.5x0.5 degree scale) measurements

* ndvi\_se – Pixel southeast of city centroid
* ndvi\_sw – Pixel southwest of city centroid
* ndvi\_ne – Pixel northeast of city centroid
* ndvi\_nw – Pixel northwest of city centroid

### Feature data example

For example, a single row in the dataset, indexed by (city, year, weekofyear): (sj, 1994, 18), has these values:

|  |  |
| --- | --- |
| **week\_start\_date** | 1994-05-07 |
| **total\_cases** | 22 |
| **station\_max\_temp\_c** | 33.3 |
| **station\_avg\_temp\_c** | 27.7571428571 |
| **station\_precip\_mm** | 10.5 |
| **station\_min\_temp\_c** | 22.8 |
| **station\_diur\_temp\_rng\_c** | 7.7 |
| **precipitation\_amt\_mm** | 68.0 |
| **reanalysis\_sat\_precip\_amt\_mm** | 68.0 |
| **reanalysis\_dew\_point\_temp\_k** | 295.235714286 |
| **reanalysis\_air\_temp\_k** | 298.927142857 |
| **reanalysis\_relative\_humidity\_percent** | 80.3528571429 |
| **reanalysis\_specific\_humidity\_g\_per\_kg** | 16.6214285714 |
| **reanalysis\_precip\_amt\_kg\_per\_m2** | 14.1 |
| **reanalysis\_max\_air\_temp\_k** | 301.1 |
| **reanalysis\_min\_air\_temp\_k** | 297.0 |
| **reanalysis\_avg\_temp\_k** | 299.092857143 |
| **reanalysis\_tdtr\_k** | 2.67142857143 |
| **ndvi\_location\_1** | 0.1644143 |
| **ndvi\_location\_2** | 0.0652 |
| **ndvi\_location\_3** | 0.1321429 |
| **ndvi\_location\_4** | 0.08175 |

## Performance metric

Performance is evaluated according to the [mean absolute error](https://en.wikipedia.org/wiki/Mean_absolute_error).

## Submission format

The format for the submission file is simply the (city, year, weekofyear) and the predicted total\_cases for San Juan or Iquitos (for an example, see SubmissionFormat.csv on the data download page). The total\_cases should be represented as integer values.

For example, if you just predicted that there were 5 cases each week for 5 weeks in San Juan and 3 cases each week for 5 weeks in Iquitos, for a total of 10 weeks, you would have the following predictions:

| **city** | **year** | **weekofyear** | **total\_cases** |
| --- | --- | --- | --- |
| **sj** | 2008 | 18 | 5 |
| **sj** | 2008 | 19 | 5 |
| **sj** | 2008 | 20 | 5 |
| **sj** | 2008 | 21 | 5 |
| **sj** | 2008 | 22 | 5 |
| **...** |  |  |  |
| **iq** | 2013 | 22 | 3 |
| **iq** | 2013 | 23 | 3 |
| **iq** | 2013 | 24 | 3 |
| **iq** | 2013 | 25 | 3 |
| **iq** | 2013 | 26 | 3 |

Your .csv file that you submit would look like:

city,year,weekofyear,total\_cases

sj,2008,18,5

sj,2008,19,5

sj,2008,20,5

sj,2008,21,5

sj,2008,22,5

...

iq,2013,22,3

iq,2013,23,3

iq,2013,24,3

iq,2013,25,3

iq,2013,26,3

**Keep in mind that you need to submit one csv with predictions for both cities**! Hence the requirement of the city column. Results will be parsed on our end and MAE scores will be given for each city's predictions.

## Good luck!

Good luck and enjoy this problem! If you have any questions you can always visit the [user forum](http://community.drivendata.org/)!