*Reporte: Breve descripción del problema de clasificación usando base de datos sobre dengue*

*Autor: Bruno Esposito - 28/10/17*

Cuál es nuestro objetivo?

Predecir si la persona padece dengue sin necesidad de realizar un test molecular. Para poder realizar esto, necesitamos observar un conjunto de características de la persona: edad, sexo, síntomas, entre otros. Nuestras predicciones serán realizadas en función a las características o atributos que las personas presenten. Queremos identificar patrones en la data que estén relacionados con la aparición de dengue.

Qué data tenemos?

Tenemos 49 atributos o características sobre 492 personas. Tenemos información sobre la edad, el sexo, el distrito de residencia, los síntomas (35 síntomas) y la prueba molecular de dengue.

Cómo lo vamos a hacer?

Queremos ser capaces de **predecir la prueba molecular de dengue sin tener que observar efectivamente el resultado de esta prueba**. Existen distintos modelos y metodologías para alcanzar esto. Uno de los modelos más populares es Naive Bayes. Es este, vamos a armar un modelo que maximice la probabilidad de tener positivo en la prueba molecular de dengue condicionado a los atributos o características que presenta la persona. El algoritmo clasificará a cada observación en (i) padece dengue y (ii) no padece dengue.

Cómo medir si estamos prediciendo bien?

Luego de implementar el modelo, hemos predicho si cada observación tiene dengue o no. Además de esto, conocemos si la persona de verdad padece dengue o no gracias al test molecular. Entonces, vamos a calcular la cantidad de veces en las que el modelo ha acertado. El modelo acierta cuando (i) predecimos padece dengue y el test molecular de la observación es positivo, y (ii) predecimos que no padece dengue y el test molecular de la observación en negativo. En la literatura, al primer grupo se le llama Verdaderos Positivos; el segundo, Verdaderos Negativos. Luego, sumamos la cantidad de aciertos (Verdaderos Positivos + Verdaderos Negativos) y lo dividimos entre el total de observaciones. Este resultado, conocido como accuracy, es el porcentaje de aciertos que tuvimos. Mientras mejor sea el accuracy, mejor poder predictivo tendrá el modelo.

Resultados preliminares

Luego de implementar cerca de diez algoritmos, **he podido predecir la presencia de dengue con un 65% de accuracy**. Esto todavía es bajo. Creo que el problema radica en que la base de datos tiene a personas que padecen otras infecciones, incluir estas categorías en la predicción podría mejorar el accuracy. Voy a probar esto durante la semana.