

# Algunas Aplicaciones del Aprendizaje de Máquinas al Diseño de Políticas Públicas

Alvaro J. Riascos Villegas  
Universidad de los Andes y Quantil

Mayo 8 de 2018

# Contenido

- 1 Hospitalizaciones Prevenibles
  - Literatura Relacionada y Datos
  - Aprendizaje de Máquinas y Teoría de la Decisión
  - Modelo Teoría de la Decisión
  - Resultados: Ahorro potencial
- 2 Ajuste de Riesgo
  - Problemas
  - Resultados
- 3 Modelos Matemáticos del Crimen
  - Validación
  - Aplicativo
- 4 Trabajo en Proceso

# Contenido

- 1 Hospitalizaciones Prevenibles
  - Literatura Relacionada y Datos
  - Aprendizaje de Máquinas y Teoría de la Decisión
  - Modelo Teoría de la Decisión
  - Resultados: Ahorro potencial
- 2 Ajuste de Riesgo
  - Problemas
  - Resultados
- 3 Modelos Matemáticos del Crimen
  - Validación
  - Aplicativo
- 4 Trabajo en Proceso

# Hospitalizaciones prevenibles

- Hospitalizaciones evitables son una fuente de incrementos en el gasto en salud.
- En Estados Unidos se estima en más USD 30 billones el costo de las hospitalizaciones innecesarias (Health Heritage Prize).
- Para darse una idea, Health Heritage Foundation ofreció 3 millones de dólares en premios al mejor modelo predictivo de hospitalizaciones.

# Hospitalizaciones prevenibles

- Hospitalizaciones evitables son una fuente de incrementos en el gasto en salud.
- En Estados Unidos se estima en más USD 30 billones el costo de las hospitalizaciones innecesarias (Health Heritage Prize).
- Para darse una idea, Health Heritage Foundation ofreció 3 millones de dólares en premios al mejor modelo predictivo de hospitalizaciones.

# Hospitalizaciones prevenibles

- Hospitalizaciones evitables son una fuente de incrementos en el gasto en salud.
- En Estados Unidos se estima en más USD 30 billones el costo de las hospitalizaciones innecesarias (Health Heritage Prize).
- Para darse una idea, Health Heritage Foundation ofreció 3 millones de dólares en premios al mejor modelo predictivo de hospitalizaciones.

- Usamos un panel de la base de datos de suficiencia del Ministerio de Salud y Protección Social entre 2009 - 2011.
- Contribuimos a la literatura con dos ideas complementarias:
  - 1 Construimos un algoritmo de aprendizaje de máquinas que predice relativamente bien los días de estadia con un año de anticipación.
  - 2 Introducimos un modelo de decisión que permite evaluar la costo efectividad de los programas de prevención de individuos con alto riesgo de hospitalización.

- Usamos un panel de la base de datos de suficiencia del Ministerio de Salud y Protección Social entre 2009 - 2011.
- Contribuimos a la literatura con dos ideas complementarias:
  - 1 Construimos un algoritmo de aprendizaje de máquinas que predice relativamente bien los días de estadia con un año de anticipación.
  - 2 Introducimos un modelo de decisión que permite evaluar la costo efectividad de los programas de prevención de individuos con alto riesgo de hospitalización.

- Usamos un panel de la base de datos de suficiencia del Ministerio de Salud y Protección Social entre 2009 - 2011.
- Contribuimos a la literatura con dos ideas complementarias:
  - 1 Construimos un algoritmo de aprendizaje de máquinas que predice relativamente bien los días de estadia con un año de anticipación.
  - 2 Introducimos un modelo de decisión que permite evaluar la costo efectividad de los programas de prevención de individuos con alto riesgo de hospitalización.

- Usamos un panel de la base de datos de suficiencia del Ministerio de Salud y Protección Social entre 2009 - 2011.
- Contribuimos a la literatura con dos ideas complementarias:
  - 1 Construimos un algoritmo de aprendizaje de máquinas que predice relativamente bien los días de estadia con un año de anticipación.
  - 2 Introducimos un modelo de decisión que permite evaluar la costo efectividad de los programas de prevención de individuos con alto riesgo de hospitalización.

## Literatura Relacionada

- 1 Chang et. al 2016. Prediction of Length of Stay of First-Ever Ischemic Stroke.
- 2 Ali, et.al. Predicting Hospital Length of Stay PHLOS: A Multi-Tiered Data Mining Approach.
- 3 Chertow. 2005. Acute Kidney Injury, Mortality, Length of Stay, and Costos in Hospitalized Patients.
- 4 Clague, et.al 2002. Predictors of outcome following hip fracture. Admission time predicts length of stay and in-hospital mortality.
- 5 Rezaei, et.al 2013. Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients.
- 6 **Bayati, et.al 2014. Data-Driven decisions for reducing readmissions for heart failure: general methodology and case study.**

- Tenemos un panel de 5.7 million afiliados al regimen contributivo entre 2009 y 2011.
- Tenemos edad, sexo, localización, IPS, EPS, diagnóstico (codificación internacional ICD10) días de estacia, costo del servicio, ingreso base de cotización, etc.
- Usamos una muestra aleatoria de un millón de individuos para entrenamiento y un millón para prueba.

- Tenemos un panel de 5.7 million afiliados al regimen contributivo entre 2009 y 2011.
- Tenemos edad, sexo, localización, IPS, EPS, diagnóstico (codificación internacional ICD10) días de estacia, costo del servicio, ingreso base de cotización, etc.
- Usamos una muestra aleatoria de un millón de individuos para entrenamiento y un millón para prueba.

- Tenemos un panel de 5.7 million afiliados al regimen contributivo entre 2009 y 2011.
- Tenemos edad, sexo, localización, IPS, EPS, diagnóstico (codificación internacional ICD10) días de estacia, costo del servicio, ingreso base de cotización, etc.
- Usamos una muestra aleatoria de un millón de individuos para entrenamiento y un millón para prueba.

- Construimos varias características con datos entre  $t - 2$  to  $t - 1$ :

Annual LOS, average LOS, maximum LOS, second maximum LOS, indicator of annual LOS greater than 30 days, standard deviation of LOS, average cost, standard deviation of cost, average income of enrollees in each insurer, standard deviation of income in each insurer, indicators of the 10 costlier diagnoses in the sample, number of hemograms, pressure tests, CTs, creatinine tests, thyroid tests, ER services, ambulatory services, hospital services, domiciliary services, drug claims, and the number of different long-term diseases affecting each patient. We also create the number of claims per month and per day of week, indicators of long-term diseases, and interactions between indicators of hospital services, ER services, domiciliary services and ambulatory services.

- Construimos varias características con datos entre  $t - 2$  to  $t - 1$ :

Annual LOS, average LOS, maximum LOS, second maximum LOS, indicator of annual LOS greater than 30 days, standard deviation of LOS, average cost, standard deviation of cost, average income of enrollees in each insurer, standard deviation of income in each insurer, indicators of the 10 costlier diagnoses in the sample, number of hemograms, pressure tests, CTs, creatinine tests, thyroid tests, ER services, ambulatory services, hospital services, domiciliary services, drug claims, and the number of different long-term diseases affecting each patient. We also create the number of claims per month and per day of week, indicators of long-term diseases, and interactions between indicators of hospital services, ER services, domiciliary services and ambulatory services.

## El Problema de Aprendizaje de Máquinas

- Tarea: Predecir los días de hospitalización en  $t$  usando la información de  $t - 1$  y  $t - 2$ . La variable dependiente es  $\ln(\text{LOS} + 1)$  en  $t$ .
- El problema es idéntico al puesto en el premio de Health Heritage Foundation.
- Como medida de rendimiento se utilizó (RMSE). El ganador del concurso obtuvo un RMSE de 0,4438 que es 2,5 veces el promedio de  $\ln(\text{LOS} + 1)$  del tercer año.

## El Problema de Aprendizaje de Máquinas

- Tarea: Predecir los días de hospitalización en  $t$  usando la información de  $t - 1$  y  $t - 2$ . La variable dependiente es  $\ln(LOS + 1)$  en  $t$ .
- El problema es idéntico al puesto en el premio de Health Heritage Foundation.
- Como medida de rendimiento se utilizó (RMSE). El ganador del concurso obtuvo un RMSE de 0,4438 que es 2,5 veces el promedio de  $\ln(LOS + 1)$  del tercer año.

## El Problema de Aprendizaje de Máquinas

- Tarea: Predecir los días de hospitalización en  $t$  usando la información de  $t - 1$  y  $t - 2$ . La variable dependiente es  $\ln(LOS + 1)$  en  $t$ .
- El problema es idéntico al puesto en el premio de Health Heritage Foundation.
- Como medida de rendimiento se utilizó (RMSE). El ganador del concurso obtuvo un RMSE de 0,4438 que es 2,5 veces el promedio de  $\ln(LOS + 1)$  del tercer año.

# Resultados

## Cuadro: Out-of-sample model fit

| Model | MAE    | RMSE   | R-squared |
|-------|--------|--------|-----------|
| OLS   | 0.4546 | 0.7502 | 0.1731    |
| ANN   | 0.5032 | 0.7824 | 0.1006    |
| RF    | 0.2634 | 0.5623 | 0.5354    |
| BT    | 0.2721 | 0.5720 | 0.5192    |
| ENS   | 0.2523 | 0.5609 | 0.5179    |

- El RMSE del ensemble lineal es 75 % el promedio de  $\ln(LOS + 1)$ , mientras que el ganador del premio Health Heritage es 249 %

# Resultados

## Cuadro: Out-of-sample model fit

| Model | MAE    | RMSE   | R-squared |
|-------|--------|--------|-----------|
| OLS   | 0.4546 | 0.7502 | 0.1731    |
| ANN   | 0.5032 | 0.7824 | 0.1006    |
| RF    | 0.2634 | 0.5623 | 0.5354    |
| BT    | 0.2721 | 0.5720 | 0.5192    |
| ENS   | 0.2523 | 0.5609 | 0.5179    |

- El RMSE del ensemble lineal es 75 % el promedio de  $\ln(LOS + 1)$ , mientras que el ganador del premio Health Heritage es 249 %

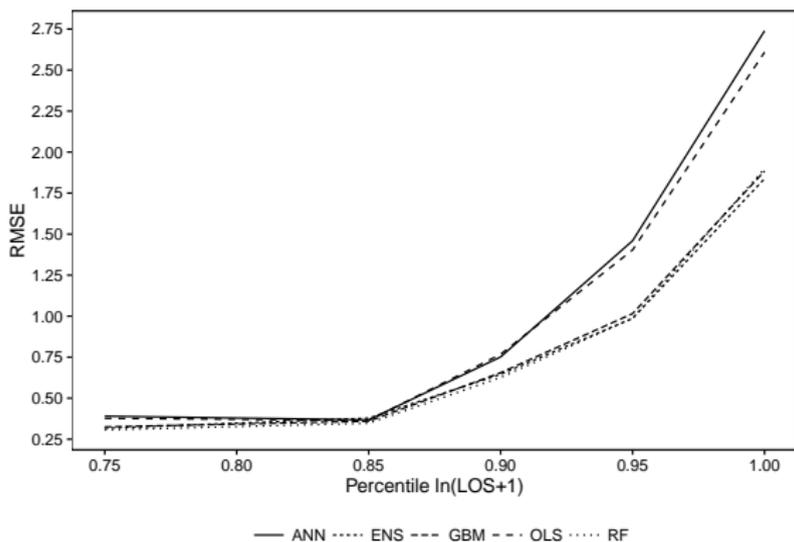


Figura: Variation in the RMSE by percentiles of the LOS distribution

# Problema Clasificación

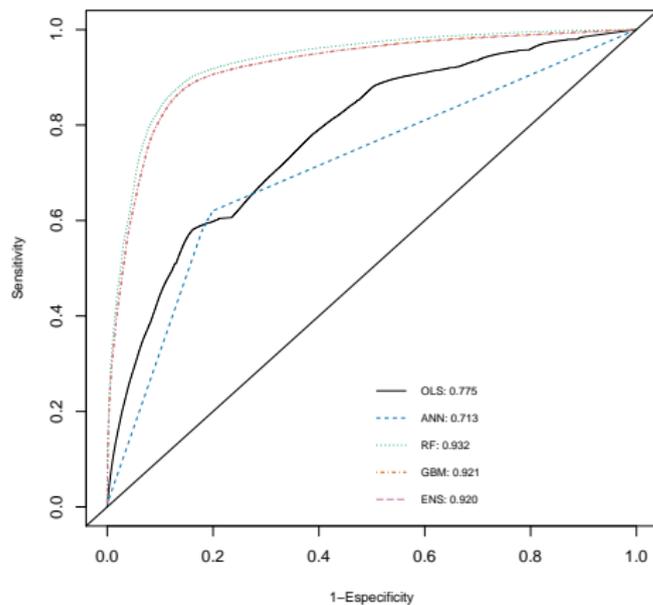


Figura: Prediction accuracy

# Resultados: Importancia relativa

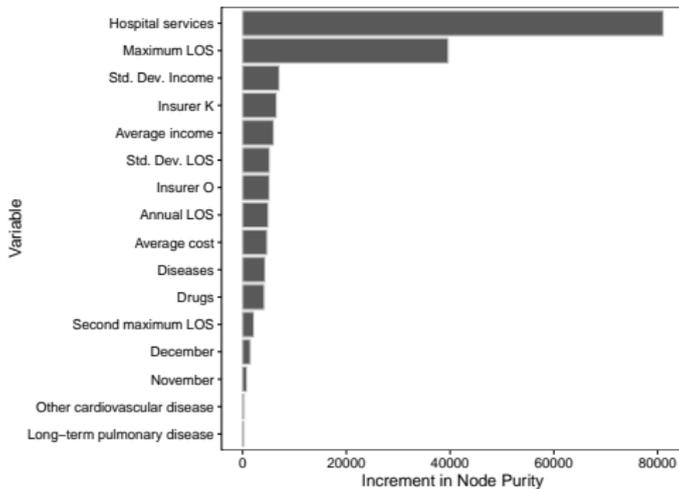


Figura: Risk factors in the random forest model

## Modelo Teoría de la Decisión

- Siguiendo a Bayati, et.al 2014 el costo esperado de hospitalizar al individuo  $i$  es el producto entre la probabilidad de ser hospitalizado y el costo de ser hospitalizado del grupo de riesgo al que pertenece:

$$C_0(\hat{p}_i) = \hat{p}_i c_g \quad (1)$$

- Si un asegurador implementa un plan de prevención la probabilidad de ser hospitalizado disminuye de acuerdo a la eficacia del programa  $\alpha$ , pero también tiene unos costos fijos  $f$ . Si el individuo es intervenido los costos esperados son:

$$C_1(\hat{p}_i) = (1 - \alpha)\hat{p}_i c_g + f \quad (2)$$

## Modelo Teoría de la Decisión

- Siguiendo a Bayati, et.al 2014 el costo esperado de hospitalizar al individuo  $i$  es el producto entre la probabilidad de ser hospitalizado y el costo de ser hospitalizado del grupo de riesgo al que pertenece:

$$C_0(\hat{p}_i) = \hat{p}_i c_g \quad (1)$$

- Si un asegurador implementa un plan de prevención la probabilidad de ser hospitalizado disminuye de acuerdo a la eficacia del programa  $\alpha$ , pero también tiene unos costos fijos  $f$ . Si el individuo es intervenido los costos esperados son:

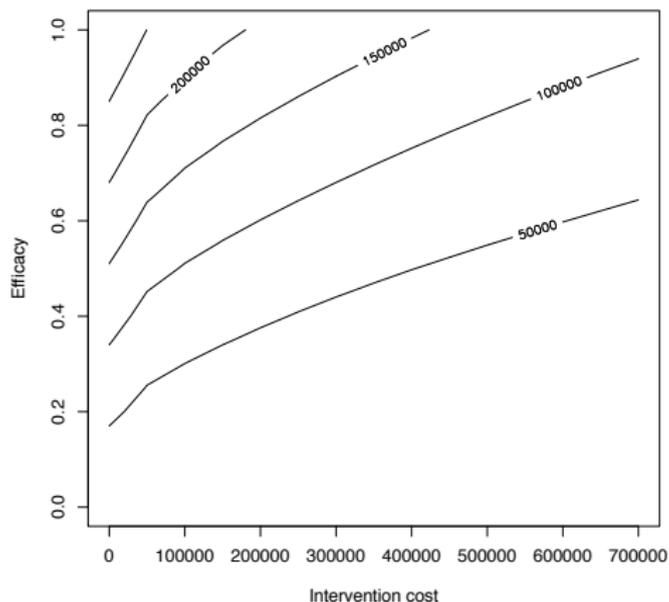
$$C_1(\hat{p}_i) = (1 - \alpha)\hat{p}_i c_g + f \quad (2)$$

- Luego, un individuo debe ser intervenido si:

$$\pi(\hat{p}_i|\alpha, f) = C_0(\hat{p}_i) - C_1(\hat{p}_i) \geq 0 \quad (3)$$

- Comparamos los costos esperados de utilizar el modelo de predicción con base la rela de decisión anterior contra dos escenarios (dados  $\alpha$  and  $f$ ):
  - 1 Política de no intervenir.
  - 2 Mejor intervención uniforme.

# Ahorro potencial contra no intervención



# Ahorro potencial contra intervención uniforme

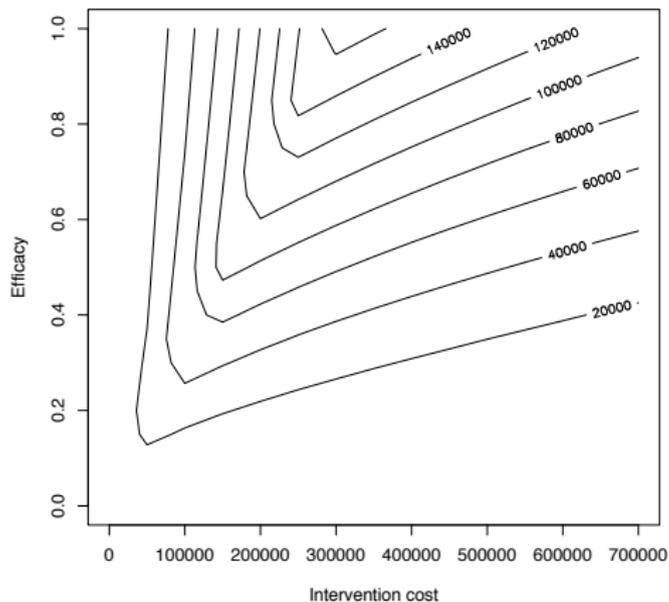


Figura: Cost savings over best uniform policy

# Contenido

- 1 Hospitalizaciones Prevenibles
  - Literatura Relacionada y Datos
  - Apredizaje de Máquinas y Teoría de la Decisión
  - Modelo Teoría de la Decisión
  - Resultados: Ahorro potencial
- 2 Ajuste de Riesgo
  - Problemas
  - Resultados
- 3 Modelos Matemáticos del Crimen
  - Validación
  - Aplicativo
- 4 Trabajo en Proceso

# Introducción

- La ley 100 de 1993 transformó el sistema colombiano de salud en un mercado de aseguramiento competitivo.
- Elementos fundamentales en la organización del mercado: POS, IPS, EPS, UPC.
- Usando un modelo de regresión lineal, el gobierno distribuye más de COP 24 billones de pesos entre las EPSs.
- Esto se hace usando como principal insumo la base de suficiencia que registra todas las atenciones, servicios, hospitalizaciones, etc. de 20 millones de Colombianos en el POS (tiene más de 450 millones de registros).

# Introducción

- La ley 100 de 1993 transformó el sistema colombiano de salud en un mercado de aseguramiento competitivo.
- Elementos fundamentales en la organización del mercado: POS, IPS, EPS, UPC.
- Usando un modelo de regresión lineal, el gobierno distribuye más de COP 24 billones de pesos entre las EPSs.
- Esto se hace usando como principal insumo la base de suficiencia que registra todas las atenciones, servicios, hospitalizaciones, etc. de 20 millones de Colombianos en el POS (tiene más de 450 millones de registros).

# Introducción

- La ley 100 de 1993 transformó el sistema colombiano de salud en un mercado de aseguramiento competitivo.
- Elementos fundamentales en la organización del mercado: POS, IPS, EPS, UPC.
- Usando un modelo de regresión lineal, el gobierno distribuye más de COP 24 billones de pesos entre las EPSs.
- Esto se hace usando como principal insumo la base de suficiencia que registra todas las atenciones, servicios, hospitalizaciones, etc. de 20 millones de Colombianos en el POS (tiene más de 450 millones de registros).

# Introducción

- La ley 100 de 1993 transformó el sistema colombiano de salud en un mercado de aseguramiento competitivo.
- Elementos fundamentales en la organización del mercado: POS, IPS, EPS, UPC.
- Usando un modelo de regresión lineal, el gobierno distribuye más de COP 24 billones de pesos entre las EPSs.
- Esto se hace usando como principal insumo la base de suficiencia que registra todas las atenciones, servicios, hospitalizaciones, etc. de 20 millones de Colombianos en el POS (tiene más de 450 millones de registros).

# Eficiencia

- Contención del gasto.
- Solución:
  - 1 Pagar con anterioridad a la prestación de los servicios (pago ex-ante).
  - 2 UPC debe reflejar el gasto esperado de salud de los afiliados.

# Eficiencia

- Contención del gasto.
- Solución:
  - 1 Pagar con anterioridad a la prestación de los servicios (pago ex-ante).
  - 2 UPC debe reflejar el gasto esperado de salud de los afiliados.

# Eficiencia

- Contención del gasto.
- Solución:
  - 1 Pagar con anterioridad a la prestación de los servicios (pago ex-ante).
  - 2 UPC debe reflejar el gasto esperado de salud de los afiliados.

# Eficiencia

- Contención del gasto.
- Solución:
  - 1 Pagar con anterioridad a la prestación de los servicios (pago ex-ante).
  - 2 UPC debe reflejar el gasto esperado de salud de los afiliados.

- Descreme del mercado mediante estrategias sutiles: calidad del servicio, largas colas, tiempos prolongados para obtener citas, etc.
- Solución:
  - 1 Ajuste de riesgo ex ante a la UPC.
  - 2 El auste de riesgo debe compensar por riesgos predecibles y socialmente aceptables.
  - 3 Mejor uso de la información.

- Descreme del mercado mediante estrategias sutiles: calidad del servicio, largas colas, tiempos prolongados para obtener citas, etc.
- Solución:
  - 1 Ajuste de riesgo ex ante a la UPC.
  - 2 El auste de riesgo debe compensar por riesgos predecibles y socialmente aceptables.
  - 3 Mejor uso de la información.

- Descreme del mercado mediante estrategias sutiles: calidad del servicio, largas colas, tiempos prolongados para obtener citas, etc.
- Solución:
  - 1 Ajuste de riesgo ex ante a la UPC.
  - 2 El auste de riesgo debe compensar por riesgos predecibles y socialmente aceptables.
  - 3 Mejor uso de la información.

- Descreme del mercado mediante estrategias sutiles: calidad del servicio, largas colas, tiempos prolongados para obtener citas, etc.
- Solución:
  - 1 Ajuste de riesgo ex ante a la UPC.
  - 2 El auste de riesgo debe compensar por riesgos predecibles y socialmente aceptables.
  - 3 Mejor uso de la información.

- Descreme del mercado mediante estrategias sutiles: calidad del servicio, largas colas, tiempos prolongados para obtener citas, etc.
- Solución:
  - 1 Ajuste de riesgo ex ante a la UPC.
  - 2 El auste de riesgo debe compensar por riesgos predecibles y socialmente aceptables.
  - 3 Mejor uso de la información.

# Capacidad predictiva del gasto

- El modelo de ajuste de riesgo del Ministerio tiene un poder predictivo normal (de acuerdo a los estándares internacionales).
- Predice el 33 % del gasto del quintil de mayor gasto de salud.
- La capacidad de predecir el gasto de ciertos **riesgos predecibles** es baja: Quedan muchos incentivos a la selección de riesgos.
- Por esta razón se hace un ajuste ex-post.

# Capacidad predictiva del gasto

- El modelo de ajuste de riesgo del Ministerio tiene un poder predictivo normal (de acuerdo a los estándares internacionales).
- Predice el 33% del gasto del quintil de mayor gasto de salud.
- La capacidad de predecir el gasto de ciertos **riesgos predecibles** es baja: Quedan muchos incentivos a la selección de riesgos.
- Por esta razón se hace un ajuste ex-post.

# Capacidad predictiva del gasto

- El modelo de ajuste de riesgo del Ministerio tiene un poder predictivo normal (de acuerdo a los estándares internacionales).
- Predice el 33% del gasto del quintil de mayor gasto de salud.
- La capacidad de predecir el gasto de ciertos **riesgos predecibles** es baja: Quedan muchos incentivos a la selección de riesgos.
- Por esta razón se hace un ajuste ex-post.

# Capacidad predictiva del gasto

- El modelo de ajuste de riesgo del Ministerio tiene un poder predictivo normal (de acuerdo a los estándares internacionales).
- Predice el 33 % del gasto del quintil de mayor gasto de salud.
- La capacidad de predecir el gasto de ciertos **riesgos predecibles** es baja: Quedan muchos incentivos a la selección de riesgos.
- Por esta razón se hace un ajuste ex-post.

# Ajuste ex-ante

**Cuadro:** Ajuste fuera de muestra distribución completa

| Modelo |              | RMSE      | MAE     | PR<br>anual | PR -<br>no anual | $R^2$ |
|--------|--------------|-----------|---------|-------------|------------------|-------|
| 1.     | WLS UPC      | 3,506,658 | 720,587 | 0.896       | 0.999            | 1.57  |
| 2.     | WLS UPC + Dx | 3,440,928 | 694,404 | 0.892       | 0.999            | 5.23  |
| 3.     | ANN FS       | 3,455,366 | 774,190 | 1.064       | 1.179            |       |
| 4.     | RF FS        | 3,465,301 | 712,820 | 0.975       | 1.087            |       |
| 5.     | GBM FS       | 3,431,044 | 721,168 | 1.002       | 1.115            |       |

# Ajuste ex-ante

**Cuadro:** Ajuste fuera de muestra en el quintil superior

| Modelo |              | RMSE      | MAE       | PR<br>anual | PR<br>no anual |
|--------|--------------|-----------|-----------|-------------|----------------|
| 1.     | WLS UPC      | 7,749,235 | 1,920,486 | 0.291       | 0.335          |
| 2.     | WLS UPC + Dx | 7,580,659 | 1,983,269 | 0.367       | 0.426          |
| 3.     | ANN FS       | 7,582,293 | 1,962,318 | 0.412       | 0.474          |
| 4.     | RF FS        | 7,580,672 | 1,988,824 | 0.424       | 0.490          |
| 5.     | GBM FS       | 7,517,520 | 1,961,026 | 0.430       | 0.500          |

# Contenido

- 1 Hospitalizaciones Prevenibles
  - Literatura Relacionada y Datos
  - Aprendizaje de Máquinas y Teoría de la Decisión
  - Modelo Teoría de la Decisión
  - Resultados: Ahorro potencial
- 2 Ajuste de Riesgo
  - Problemas
  - Resultados
- 3 Modelos Matemáticos del Crimen
  - Validación
  - Aplicativo
- 4 Trabajo en Proceso

## Introducción

- Esta literatura a comenzado a influenciar la administración de los recursos policiales de los principales centros urbanos: Los Ángeles CA, Atlanta GA, Chicago IL, New York NW, Alhambra CA, San Francisco CA, Modesto CA, Santa Cruz, CA.
- Se diseñó una metodología para comparar diferentes modelos de predicción del crimen en Bogotá.
- Esta se extiende de forma natural a otros centros urbanos con la información adecuada (Cali, Medellín).
- Datos utilizados: 329,793 crímenes ocurridos en Bogotá entre 2004 y 2014 (geolocalizados, con fecha y hora).
- Se compararon modelos de Puntos, Elipses, KDE y varias versiones de modelos espacio-temporales.

## Introducción

- Esta literatura a comenzado a influenciar la administración de los recursos policiales de los principales centros urbanos: Los Ángeles CA, Atlanta GA, Chicago IL, New York NW, Alhambra CA, San Francisco CA, Modesto CA, Santa Cruz, CA.
- Se diseñó una metodología para comparar diferentes modelos de predicción del crimen en Bogotá.
- Esta se extiende de forma natural a otros centros urbanos con la información adecuada (Cali, Medellín).
- Datos utilizados: 329,793 crímenes ocurridos en Bogotá entre 2004 y 2014 (geolocalizados, con fecha y hora).
- Se compararon modelos de Puntos, Elipses, KDE y varias versiones de modelos espacio-temporales.

## Introducción

- Esta literatura a comenzado a influenciar la administración de los recursos policiales de los principales centros urbanos: Los Ángeles CA, Atlanta GA, Chicago IL, New York NW, Alhambra CA, San Francisco CA, Modesto CA, Santa Cruz, CA.
- Se diseñó una metodología para comparar diferentes modelos de predicción del crimen en Bogotá.
- Esta se extiende de forma natural a otros centros urbanos con la información adecuada (Cali, Medellín).
- Datos utilizados: 329,793 crímenes ocurridos en Bogotá entre 2004 y 2014 (geolocalizados, con fecha y hora).
- Se compararon modelos de Puntos, Elipses, KDE y varias versiones de modelos espacio-temporales.

## Introducción

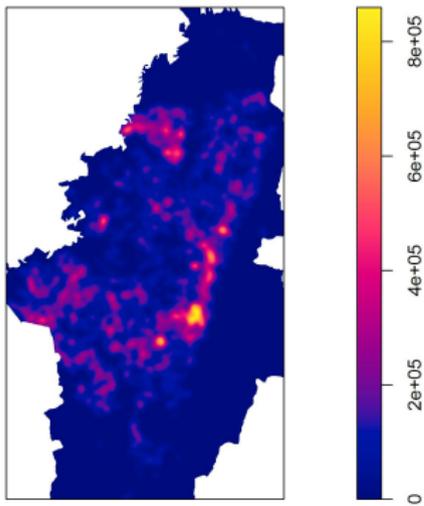
- Esta literatura a comenzado a influenciar la administración de los recursos policiales de los principales centros urbanos: Los Ángeles CA, Atlanta GA, Chicago IL, New York NW, Alhambra CA, San Francisco CA, Modesto CA, Santa Cruz, CA.
- Se diseñó una metodología para comparar diferentes modelos de predicción del crimen en Bogotá.
- Esta se extiende de forma natural a otros centros urbanos con la información adecuada (Cali, Medellín).
- Datos utilizados: 329,793 crímenes ocurridos en Bogotá entre 2004 y 2014 (geolocalizados, con fecha y hora).
- Se compararon modelos de Puntos, Elipses, KDE y varias versiones de modelos espacio-temporales.

## Introducción

- Esta literatura a comenzado a influenciar la administración de los recursos policiales de los principales centros urbanos: Los Ángeles CA, Atlanta GA, Chicago IL, New York NW, Alhambra CA, San Francisco CA, Modesto CA, Santa Cruz, CA.
- Se diseñó una metodología para comparar diferentes modelos de predicción del crimen en Bogotá.
- Esta se extiende de forma natural a otros centros urbanos con la información adecuada (Cali, Medellín).
- Datos utilizados: 329,793 crímenes ocurridos en Bogotá entre 2004 y 2014 (geolocalizados, con fecha y hora).
- Se compararon modelos de Puntos, Elipses, KDE y varias versiones de modelos espacio-temporales.

# Modelo KDE: Ejemplo

Heat map for a 1-year period of crime



- Este es un modelo basado en la clasificación de los eventos como antecedentes y réplicas.
- Es el estado del arte en modelos de predicción del crimen.

# Modelo Espacio - Temporal: Motivación

Mohler et al.: Self-Exciting Point Process Modeling of Crime

101

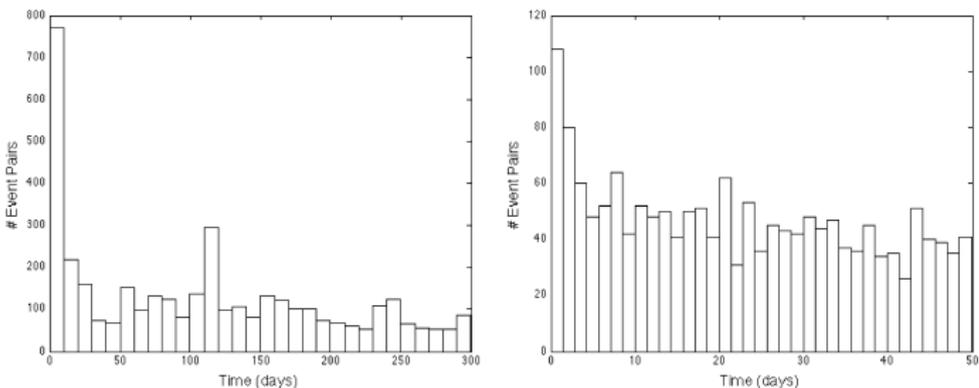


Figure 1. On the left, histogram of times (less than 300 days) between Southern California earthquake events of magnitude 3.0 or greater separated by 110 kilometers or less. On the right, histogram of times (less than 50 days) between burglary events separated by 200 meters or less.

# Modelo Espacio - Temporal: Motivación

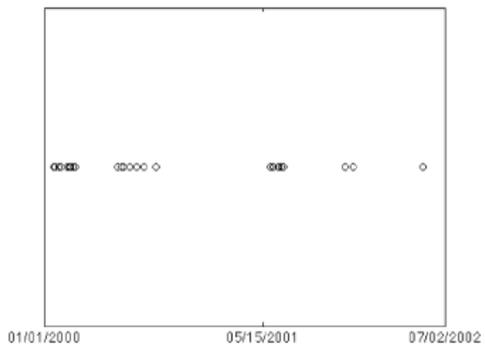


Figure 2. Times of violent crimes between two rivalry gangs in Los Angeles.

- Se considera un modelo de la intensidad espacio temporal del crimen de la forma:

$$\lambda(t, x, y) = \mu(t, x, y) + \sum_{k:t_k < t} g(t - t_k, x - x_k, y - y_k) \quad (4)$$

# Validación

- Evaluar la capacidad predictiva del modelo de predicción de crimen propuesto.
- En particular, se compara el poder predictivo del modelo al ser entrenado con diferentes horizontes de tiempo.

- Se cuenta con datos de crímenes en Bogotá del 16 de abril al 30 de junio de 2017: 16.402 datos.
- Para esta validación se usaron datos de la localidad de Santa Fé, en donde se aplicará el piloto, delimitada por latitudes entre  $[4,571, 4,629]$  y longitudes  $> -74,091$ .

- Se cuenta con datos de crímenes en Bogotá del 16 de abril al 30 de junio de 2017: 16.402 datos.
- Para esta validación se usaron datos de la localidad de Santa Fé, en donde se aplicará el piloto, delimitada por latitudes entre  $[4,571, 4,629]$  y longitudes  $> -74,091$ .

- 1.676 ( $\approx 10\%$ ) crímenes en la localidad de Santa Fé en el período tratado.
- Se entrenó el modelo de crimen con datos entre 1 y 7 semanas y se validó con las 3 semanas posteriores al entrenamiento, en todos los casos, del 10 al 30 de junio de 2017 (407 crímenes).

- 1.676 ( $\approx 10\%$ ) crímenes en la localidad de Santa Fé en el período tratado.
- Se entrenó el modelo de crimen con datos entre 1 y 7 semanas y se validó con las 3 semanas posteriores al entrenamiento, en todos los casos, del 10 al 30 de junio de 2017 (407 crímenes).

- Utilizamos el *Precision Accuracy Index*

$$\text{PAI} = \frac{\text{Hit Rate}}{\text{Percentage of Area}}$$
$$\text{Hit Rate} = \frac{\text{Crimes predicted in Hotspots}}{\text{Total Crimes}}$$
$$\text{Percentage Area} = \frac{\text{Area of Hotspots}}{\text{Total Area}}$$

- Sin embargo en modelos muy granulares no es una buena medida.

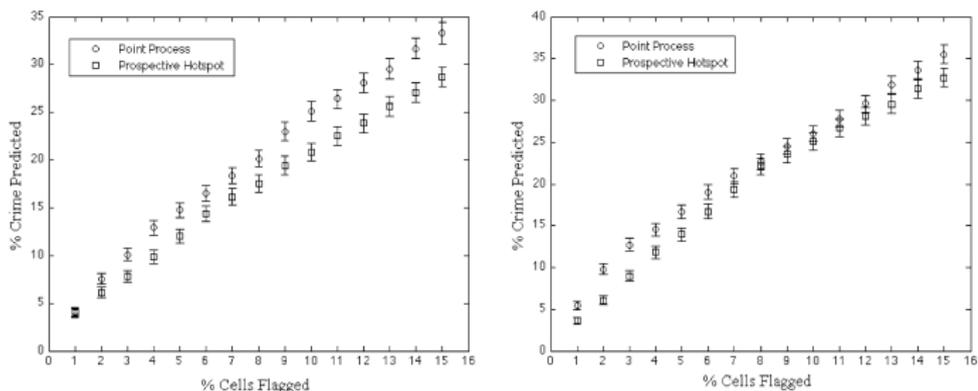


Figure 6. Forecasting strategy comparison. Average daily percentage of crimes predicted plotted against percentage of cells flagged for 2005 burglary using 200 m by 200 m cells. Error bars correspond to the standard error. Prospective hotspot cutoff parameters are 400 meters and 8 weeks (left) and optimal parameters (right) are 200 meters and 39 weeks. Spatial background intensity  $\mu(x, y)$  smoothing bandwidth for the point process is 300 meters (left) selected by cross validation and 130 meters (right) selected to optimize the number of crimes predicted.

- Se divide Bogotá ( $1,547\text{km}^2$ ) en 10,946 celdas de  $\approx 145\text{m}^2$  cada una; 1,019 celdas en la localidad de Santa Fé.
- Se entrena el modelo con el correspondiente número de semanas y se predicen los puntos calientes para cada turno de 8 horas, definidos como el 10% de las celdas con mayor probabilidad de crimen.
- Se investigan cuántos crímenes de los datos de validación ocurrieron en los puntos calientes predichos por el modelo.

- Se divide Bogotá ( $1,547\text{km}^2$ ) en 10,946 celdas de  $\approx 145\text{m}^2$  cada una; 1,019 celdas en la localidad de Santa Fé.
- Se entrena el modelo con el correspondiente número de semanas y se predicen los puntos calientes para cada turno de 8 horas, definidos como el 10% de las celdas con mayor probabilidad de crimen.
- Se investigan cuántos crímenes de los datos de validación ocurrieron en los puntos calientes predichos por el modelo.

- Se divide Bogotá ( $1,547\text{km}^2$ ) en 10,946 celdas de  $\approx 145\text{m}^2$  cada una; 1,019 celdas en la localidad de Santa Fé.
- Se entrena el modelo con el correspondiente número de semanas y se predicen los puntos calientes para cada turno de 8 horas, definidos como el 10% de las celdas con mayor probabilidad de crimen.
- Se investigan cuántos crímenes de los datos de validación ocurrieron en los puntos calientes predichos por el modelo.

HIT RATE PROMEDIO POR SEMANAS DE ENTRENAMIENTO



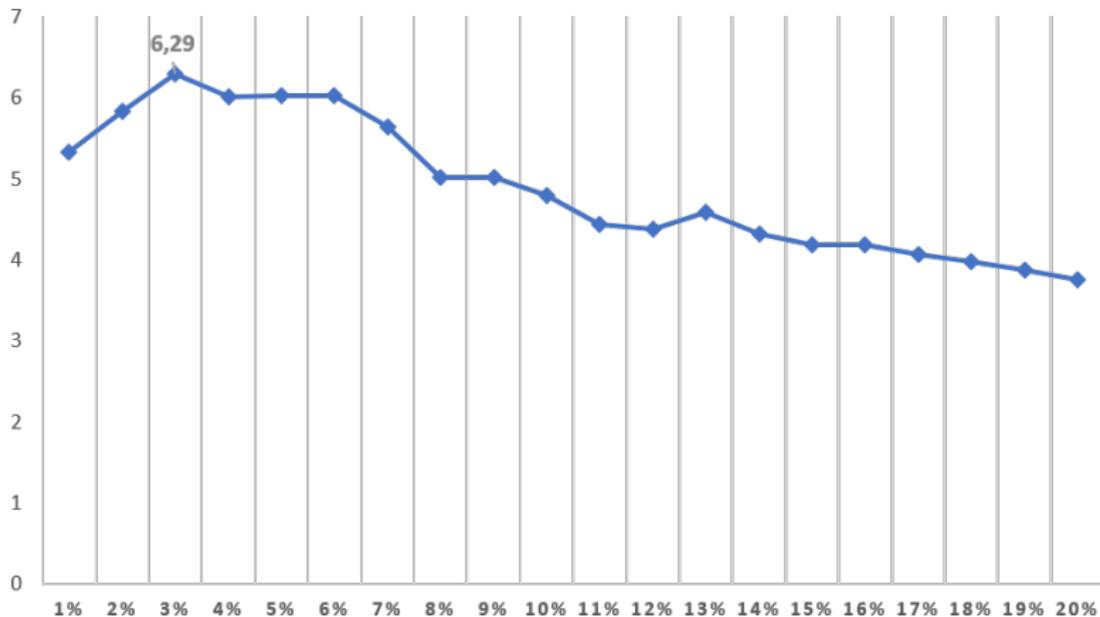
- Hit Rate con 7 semanas de entrenamiento y 10% de cobertura de puntos calientes:

| Predicción | bw fijo | bw variable | KDE  |
|------------|---------|-------------|------|
| Semana 1   | 0,44    | 0,57        | 0,42 |
| Semana 2   | 0,46    | 0,59        | 0,44 |
| Semana 3   | 0,54    | 0,62        | 0,53 |
| Promedio   | 0,48    | 0,59        | 0,46 |

- Hit Rate con 7 semanas de entrenamiento y 10% de cobertura de puntos calientes:

| Predicción | bw fijo | bw variable | KDE  |
|------------|---------|-------------|------|
| Semana 1   | 0,44    | 0,57        | 0,42 |
| Semana 2   | 0,46    | 0,59        | 0,44 |
| Semana 3   | 0,54    | 0,62        | 0,53 |
| Promedio   | 0,48    | 0,59        | 0,46 |

## PAI PROMEDIO SEGÚN PORCENTAJE DE COBERTURA



~/R/Code/codigo\_crimen/App - Shiny

http://127.0.0.1:5084 Open in Browser Publish

### Prediccion de crimen

quartil matemáticas aplicadas

Horas: 10

Día:  
 lunes  
 martes  
 miercoles  
 jueves  
 viernes  
 sabado  
 domingo

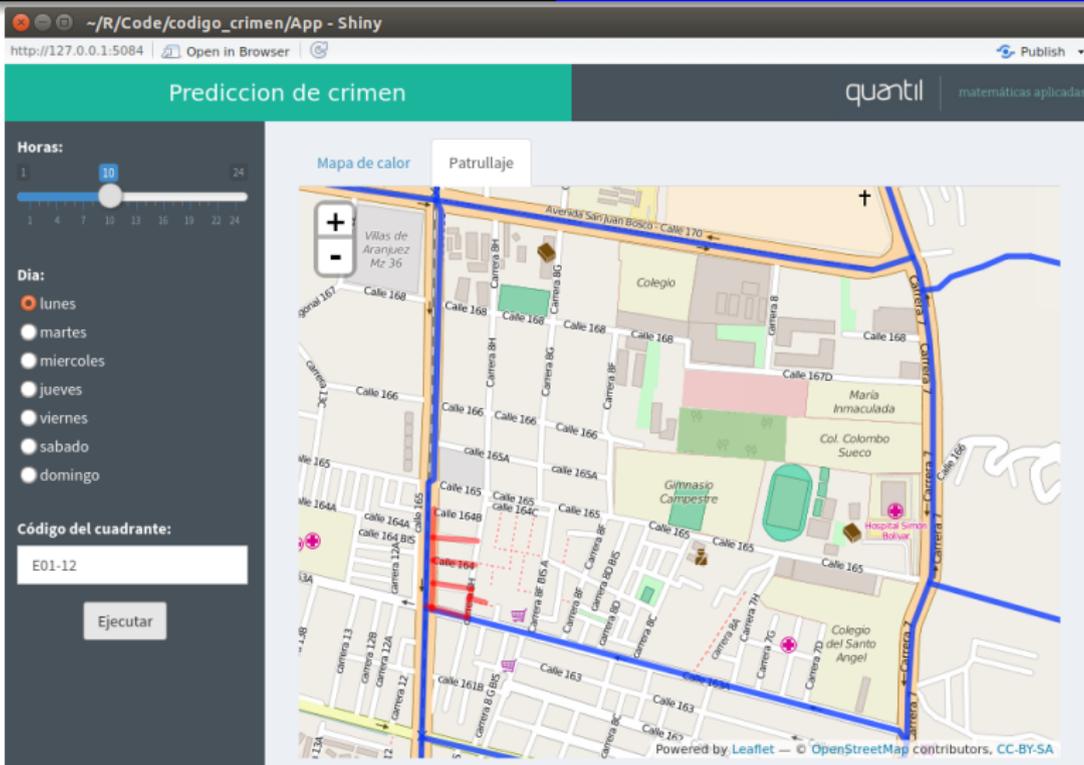
Código del cuadrante:  
E01-13

Ejecutar

Mapa de calor Patruillaje

Powered by Leaflet — © OpenStreetMap contributors, CC-BY-SA

## Patruillaje por segmentos



## Patrullaje por segmentos

# Contenido

- 1 Hospitalizaciones Prevenibles
  - Literatura Relacionada y Datos
  - Apredizaje de Máquinas y Teoría de la Decisión
  - Modelo Teoría de la Decisión
  - Resultados: Ahorro potencial
- 2 Ajuste de Riesgo
  - Problemas
  - Resultados
- 3 Modelos Matemáticos del Crimen
  - Validación
  - Aplicativo
- 4 Trabajo en Proceso

- Salud pública:
  - 1 Grupos de enfermedades usando análisis supervisado: Aprendizaje de máquinas, distancia entre particiones y optimización usando el algoritmo Metropolis Hasting.
  - 2 Grupos de enfermedades usando análisis no supervisado: LDA.
- Predicción del crimen (en conjunto con el Profesor Francisco Gomez).
  - 1 Teoría de las ventanas rotas, imágenes y aprendizaje profundo.
  - 2 Viajes de Levy.
  - 3 Kernel aprendizaje en variedades.

- Salud pública:
  - 1 Grupos de enfermedades usando análisis supervisado: Aprendizaje de máquinas, distancia entre particiones y optimización usando el algoritmo Metropolis Hasting.
  - 2 Grupos de enfermedades usando análisis no supervisado: LDA.
- Predicción del crimen (en conjunto con el Profesor Francisco Gomez).
  - 1 Teoría de las ventanas rotas, imágenes y aprendizaje profundo.
  - 2 Viajes de Levy.
  - 3 Kernel aprendizaje en variedades.

- Salud pública:
  - 1 Grupos de enfermedades usando análisis supervisado: Aprendizaje de máquinas, distancia entre particiones y optimización usando el algoritmo Metropolis Hasting.
  - 2 Grupos de enfermedades usando análisis no supervisado: LDA.
- Predicción del crimen (en conjunto con el Profesor Francisco Gomez).
  - 1 Teoría de las ventanas rotas, imágenes y aprendizaje profundo.
  - 2 Viajes de Levy.
  - 3 Kernel aprendizaje en variedades.

- Salud pública:
  - 1 Grupos de enfermedades usando análisis supervisado: Aprendizaje de máquinas, distancia entre particiones y optimización usando el algoritmo Metropolis Hasting.
  - 2 Grupos de enfermedades usando análisis no supervisado: LDA.
- Predicción del crimen (en conjunto con el Profesor Francisco Gomez).
  - 1 Teoría de las ventanas rotas, imágenes y aprendizaje profundo.
  - 2 Viajes de Levy.
  - 3 Kernel aprendizaje en variedades.