

Data Mining Footprints in Academics and Industry in Colombia A Personal Perspective

Alvaro J. Riascos Villegas
University of los Andes and Quantil

Leveraging Big Data for Sustainable Development
Bogotá

November 29, 2016

Contents

- 1 Introduction
- 2 Crime Prediction
- 3 Satellite Images Illegal Mining Detection
- 4 Mining Medical Data
- 5 Text Mining
- 6 Data Driven Marketing
- 7 Scoring and Classification
- 8 Public Policy

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

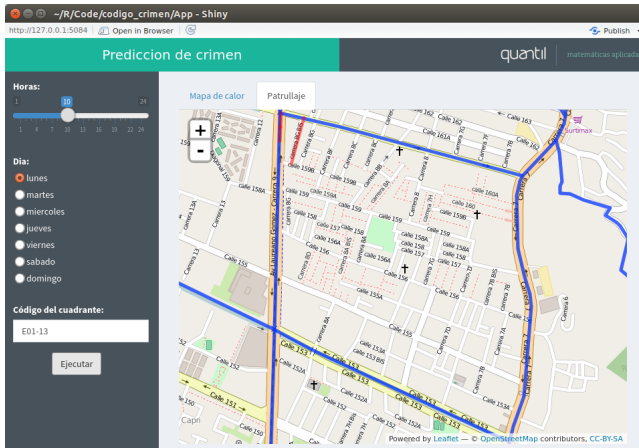
Introduction

- I provide the most significant examples of successful data mining application in academics and industry in Colombia: My very personal experience in los Andes and Quantil.
- Focus: Capacities.
- These are applications to:
 - 1 Crime prediction.
 - 2 Satellite images illegal mining detection.
 - 3 Mining medical data.
 - 4 Text mining.
 - 5 Data driven marketing.
 - 6 Scoring and classification.
 - 7 Public policy.

Contents

- 1 Introduction
- 2 Crime Prediction**
- 3 Satellite Images Illegal Mining Detection
- 4 Mining Medical Data
- 5 Text Mining
- 6 Data Driven Marketing
- 7 Scoring and Classification
- 8 Public Policy

Crime Prediction

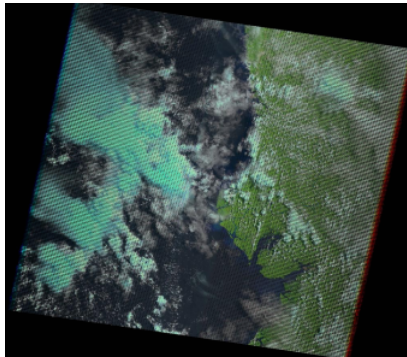


Contents

- 1 Introduction
- 2 Crime Prediction
- 3 Satellite Images Illegal Mining Detection**
- 4 Mining Medical Data
- 5 Text Mining
- 6 Data Driven Marketing
- 7 Scoring and Classification
- 8 Public Policy

Satellite Image Illegal Mining Detection

- Courtesy of Santiago Saavedra (PhD Candidate Stanford University)

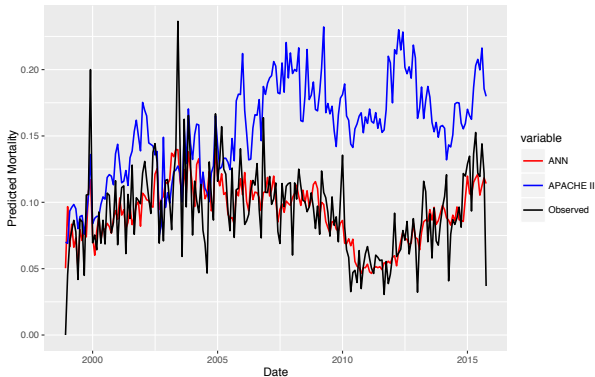


Contents

- 1 Introduction
- 2 Crime Prediction
- 3 Satellite Images Illegal Mining Detection
- 4 Mining Medical Data**
- 5 Text Mining
- 6 Data Driven Marketing
- 7 Scoring and Classification
- 8 Public Policy

Intense care unit data

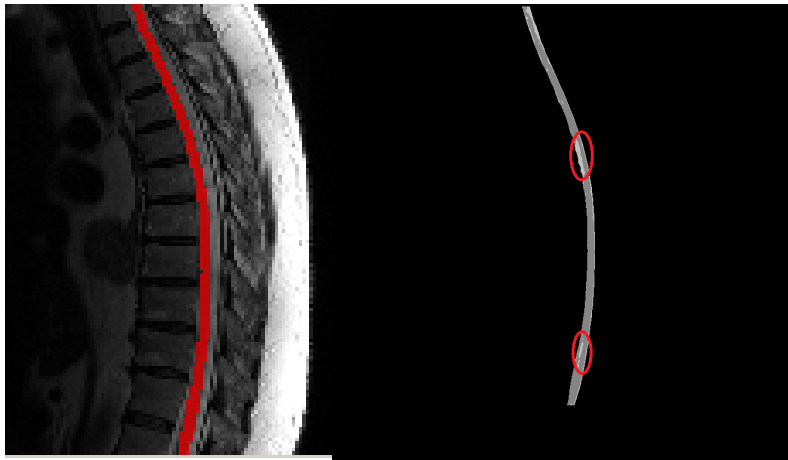
- Courtesy Fundación Valle del Lili (FVL).



Gráfica: Courtesy of Johns Hopkins University

(a) Spinal Cord

(b) Spinal Cord Injury

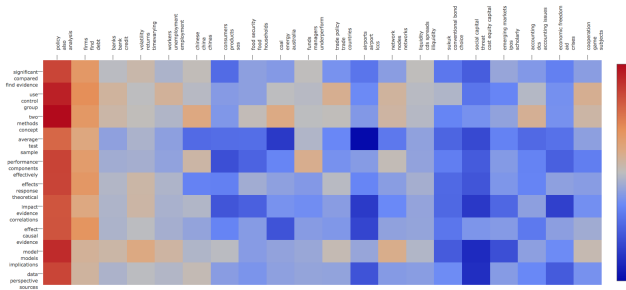


Contents

- 1 Introduction
- 2 Crime Prediction
- 3 Satellite Images Illegal Mining Detection
- 4 Mining Medical Data
- 5 Text Mining**
- 6 Data Driven Marketing
- 7 Scoring and Classification
- 8 Public Policy

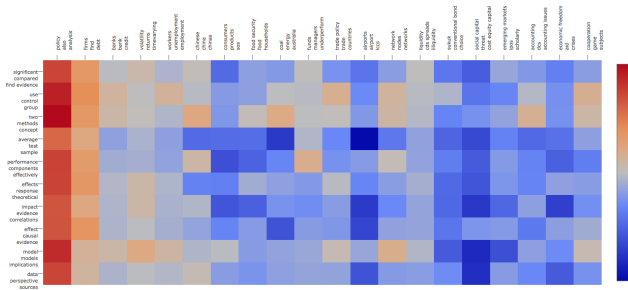
GAP MAPS

- Manual gap maps <http://gapmaps.3ieimpact.org/>
- Automatic gap maps (work in progress CESED - UNIANDES: crime and drugs gap maps).



GAP MAPS

- Manual gap maps <http://gapmaps.3ieimpact.org/>
- Automatic gap maps (work in progress CESED - UNIANDES: crime and drugs gap maps).



Automatic summarization

- Courtesy of Quantil (independent internal project).



The screenshot shows a web browser window displaying the website 'www.acuerdosdepaz.co'. The page features a dark background with a photograph of a group of people seated at a long table, likely during a press conference or meeting. The text on the page is white and centered.

quantil

HOME WORDCLOUDS RESUMEN DISTRIBUCIÓN ¿QUÉ NOS SOMOS? ACUERDO

LOGIN f t

ACUERDOS DE PAZ

El 23 de agosto de 2016 el gobierno de Colombia y las FARC firmaron el "Acuerdo para la terminación del conflicto y la construcción de una paz estable y duradera". Desde hoy hasta el 2 de Octubre de 2016 los colombianos deben revisar el texto, analizar y decidir si legitiman esta decisión.

Investigadores y matemáticos de [Quantil | Matemáticas Aplicadas](#) se pusieron en la tarea de analizar el texto algorítmicamente, usando técnicas de inteligencia artificial y procesamiento de lenguaje natural, para facilitar su entendimiento.

Cuenta regresiva para el plebiscito
DÍA DE VOTACIÓN

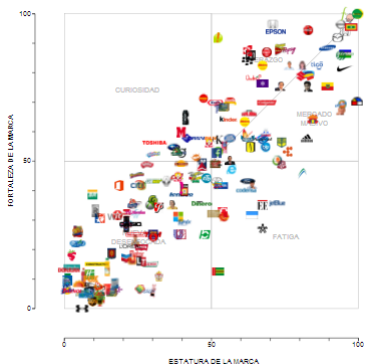
12:30 p. m.
20/10/2016

Contents

- 1 Introduction
- 2 Crime Prediction
- 3 Satellite Images Illegal Mining Detection
- 4 Mining Medical Data
- 5 Text Mining
- 6 Data Driven Marketing**
- 7 Scoring and Classification
- 8 Public Policy

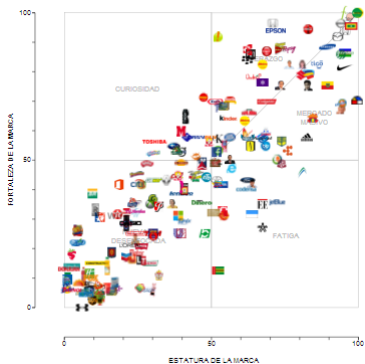
Data Driven Marketing

- Courtesy: Young & Rubican.
- Mining of social networks data (twitter).



Data Driven Marketing

- Courtesy: Young & Rubican.
- Mining of social networks data (twitter).

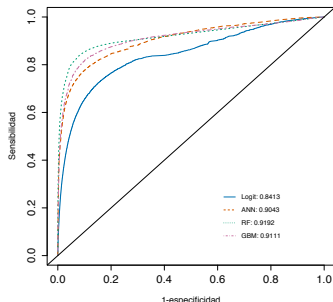


Contents

- 1 Introduction
- 2 Crime Prediction
- 3 Satellite Images Illegal Mining Detection
- 4 Mining Medical Data
- 5 Text Mining
- 6 Data Driven Marketing
- 7 Scoring and Classification**
- 8 Public Policy

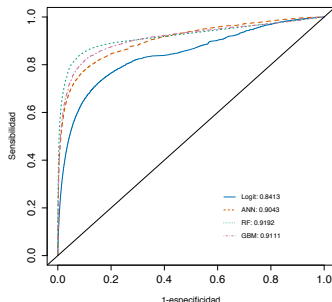
Scoring and Classification

- Cross industry applications: credit, cross selling, client retention, etc.
- Beating to death the logistic model.



Scoring and Classification

- Cross industry applications: credit, cross selling, client retention, etc.
- Beating to death the logistic model.



Contents

- 1 Introduction
- 2 Crime Prediction
- 3 Satellite Images Illegal Mining Detection
- 4 Mining Medical Data
- 5 Text Mining
- 6 Data Driven Marketing
- 7 Scoring and Classification
- 8 Public Policy**

Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Health records: 400 million data items per year, 20 million individuals.
- Potential problems: Risk adjustment, solvency, fraud, data manipulation, misreport, errors, etc.

Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Health records: 400 million data items per year, 20 million individuals.
- Potential problems: Risk adjustment, solvency, fraud, data manipulation, misreport, errors, etc.

Health Records

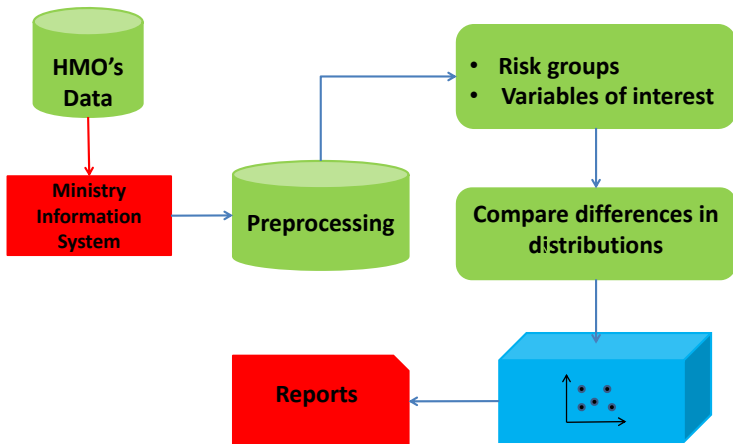
- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Health records: 400 million data items per year, 20 million individuals.
- Potential problems: Risk adjustment, solvency, fraud, data manipulation, misreport, errors, etc.

Health Records

- Colombia has a mandatory competitive health insurance system.
- Payments are determined using information provided by HMOs.
- Health records: 400 million data items per year, 20 million individuals.
- Potential problems: Risk adjustment, solvency, fraud, data manipulation, misreport, errors, etc.

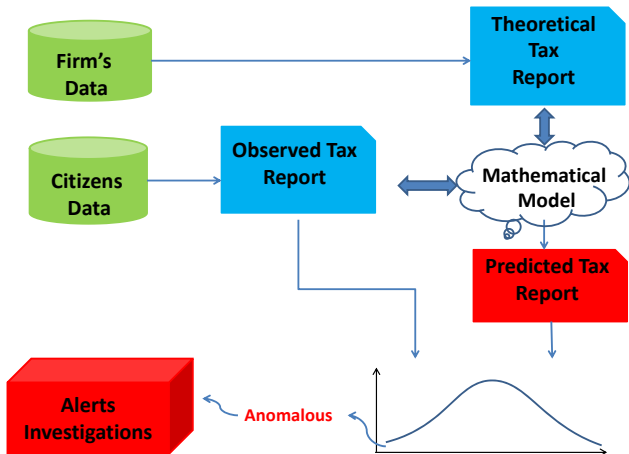
Health Records: Anomaly detection

- Courtesy of Ministry of Health and Social Protection.



Tax Evasion

- Problems: Detect likely cases of tax evasion.



Money Laundry

- Problem: Detect anomalous transactions that may be suspicious of money laundry.
- Confidentiality agreements forbids me of explaining further details.

Money Laundry

- Problem: Detect anomalous transactions that may be suspicious of money laundry.
- Confidentiality agreements forbids me of explaining further details.

Lawsuits Against the State

- Courtesy of National Agency for Judicial State Defence (ANDJE).

The screenshot displays the website of the National Agency for Judicial State Defence (ANDJE). The main heading is "ANDJE Probabilidad de que se otorgue una demanda contra el Estado". The interface includes a sidebar with various filters and a main content area showing the results of a simulation.

Filters (Left Sidebar):

- Resultado del primer fallo: Otro
- Subtipo 1: Falta judicial
- Se concede la apelación: Otro
- Subtipo 2: Penson
- Nombre del Magistrado del tribunal: Otro
- Nombre del juez: Otro
- Numero del juzgado: 4
- El demandante es: Civil
- El demandado presenta apelación: No
- El demandado cambia de apoderado: No
- El demandante cambia de apoderado: No
- Numero del tribunal: Otro
- Entidad demandada: Caja de Retiro Militar
- Tiempo entre la admision demanda y la primera sentencia, en dias: 100
- Tiempo entre los hechos y la admision de la demanda, en dias: 200
- Duracion total del proceso, en dias: 1800
- Numero de pruebas presentadas por el demandante: 10

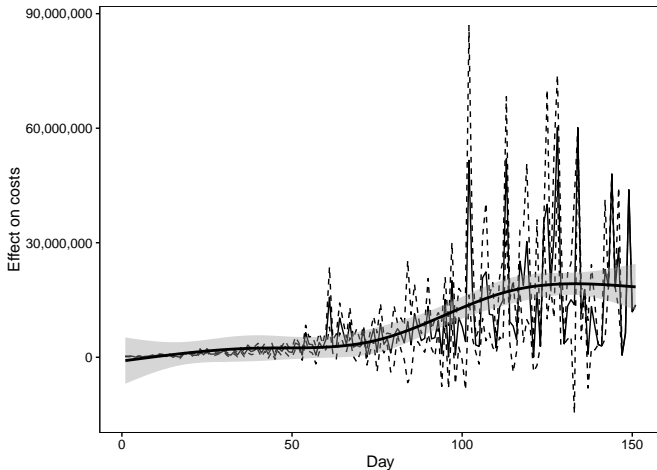
Main Content Area:

- Comienzo del proceso: Durante el proceso
- Probabilidad ajustada al comienzo del proceso: 86.2 %
- Logo of the Agencia Nacional de Defensa Juridica del Estado

The bottom of the image shows a Windows taskbar with several open applications, including a browser, Word, Excel, and PowerPoint.

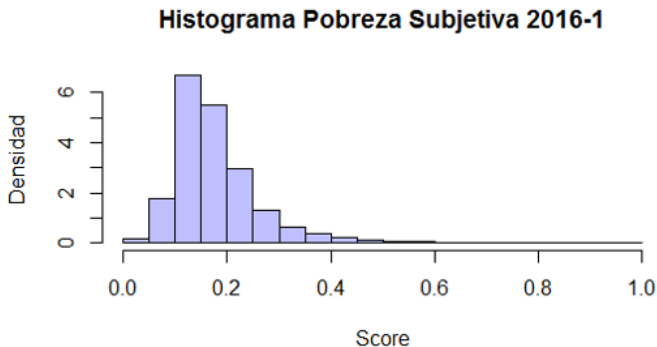
Predicting unnecessary hospitalizations

- Courtesy of Ministry of Health and Social Protection.



Subjective Poverty

- Courtesy of National Department of Statistics (DANE).



Thank you very much!

alvaro.riascos@quantil.com.co