Universidad de los Andes

Inteligencia Artificial para Predecir Corrupción en la Administración Pública Municipal de Colombia

Kevin Steven Mojica Muñoz

Tesis para optar por el título de: Magister en Economía y Magister en Políticas Públicas

Asesores: Andrés Ham y Alvaro Riascos

Objetivo de Investigación:

Generar un **indicador de riesgo de corrupción municipal** basado en predicciones futuras de variables relacionadas al fenómeno utilizando Aprendizaje de Máquinas (IA)

Impacto en términos de Política Pública:
La investigación pretende generar alertas tempranas de riesgo de corrupción en administraciones municipales que sirvan a los organismos de control para focalizar sus recursos de investigación y prevención.

Motivación:

- La corrupción tiene un impacto negativo y significativo en el crecimiento económico (Ugur, 2013)

- Se estima que el costo de la corrupción es aproximadamente el 5% del PIB mundial, lo que equivale a 2.6 trillones de dólares (Wickberg, 2013). En Colombia esta cifra puede alcanzar 4% del PIB nacional (Secretaría de Transparencia, 2020).

- El mayor riesgo de corrupción está en la administración municipal (Zuleta et al., 2018).

- Los algoritmos basados en modelos de aprendizaje de máquinas han demostrado ser herramientas útiles de predicción (Wainer, 2016; Olson et al., 2018; Chul Lee et al., 2018; Carvalho et al., 2014), pero poco se han implementado en el mundo para temas de corrupción (Aarvik, 2019).

- Hasta el momento en Colombia no se han desarrollado predicciones de corrupción basadas en algoritmos de Aprendizaje Automático a nivel municipal.

Hipótesis:

- Los algoritmos de aprendizaje máquinas logran predecir las variables relacionadas a la corrupción con un error (RMSE) cercano a una desviación estándar.

- Los algoritmos de aprendizaje máquinas tienen mejor desempeño que técnicas tradicionales de predicción: regresión lineal.

- Los indicadores basados en algoritmos de aprendizaje de máquinas logran mejor desempeño al detectar riesgo de corrupción que indicadores tradicionales.

Marco teórico:

+La probabilidad de que se cometa un acto de corrupción en un municipio depende de variables observables y no observables.

- *Hay patrones generalizados en estas variables que se mantienen en el tiempo y que permiten identificar la probabilidad futura de que se cometa un acto de corrupción.

- *La función generadora de datos es algo intangible a lo que se puede aproximar por medio de algoritmos de aprendizaje de máquinas.

$$\tilde{Y}_{t,i} = f(X_{t-1,i}, X_{t-2,i}, \ldots, X_{t-k,i}; \tilde{Y}_{t-1,i}, \tilde{Y}_{t-2,i}, \ldots, \tilde{Y}_{t-k,i}, e_{t,i})$$

$$\tilde{Y}_{t,i} = Y_{t,i} + u_{t,i}$$

Cada posible variable observable es una medida aproximada de la corrupción real del municipio

Datos:

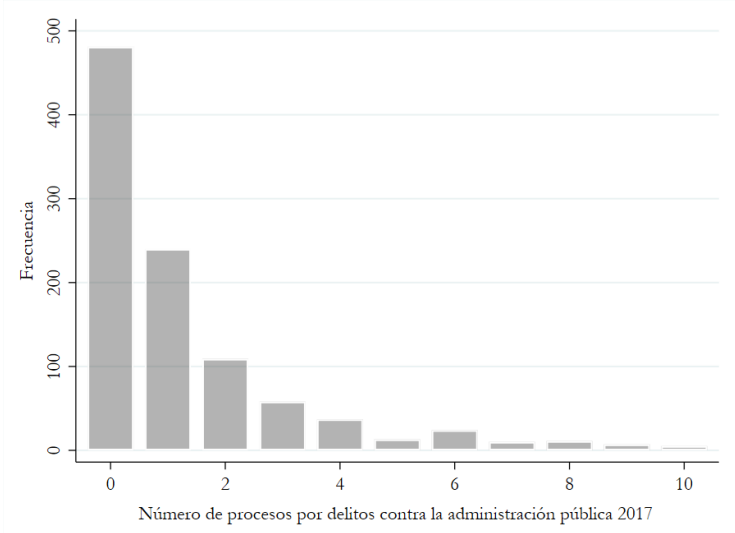
Tabla 1. Descripción de las variables objetivo

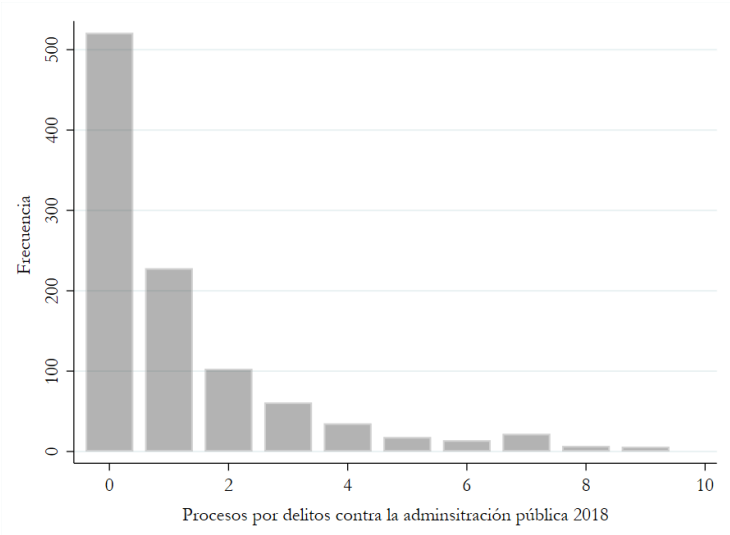
Variable		Descripción	Periodo	Fuente
Procesos	Estimación Periodo Electoral	Promedio anual de Procesos Penales por delitos contra la administración pública en el periodo electoral del alcalde.	Train: 2012-2015 Test: 2016-2019	Base de datos del Sistema Penal Oral Acusatorio (SPOA)
	Estimación Anual	Número de Procesos Penales por delitos contra la administración pública en el año en cuestión.	Train: 2017 Test: 2018	
Denuncias	Estimación Periodo Electoral	Promedio anual de denuncias por presuntos actos de corrupción en la administración municipal.	Train: 2012-2015 Test: 2016-2019	Observatorio Anticorrupción (Secretaría de Transparencia)
	Estimación Anual	Número de denuncias por presuntos actos de corrupción en la administración municipal.	Train: 2017 Test: 2018	

Train

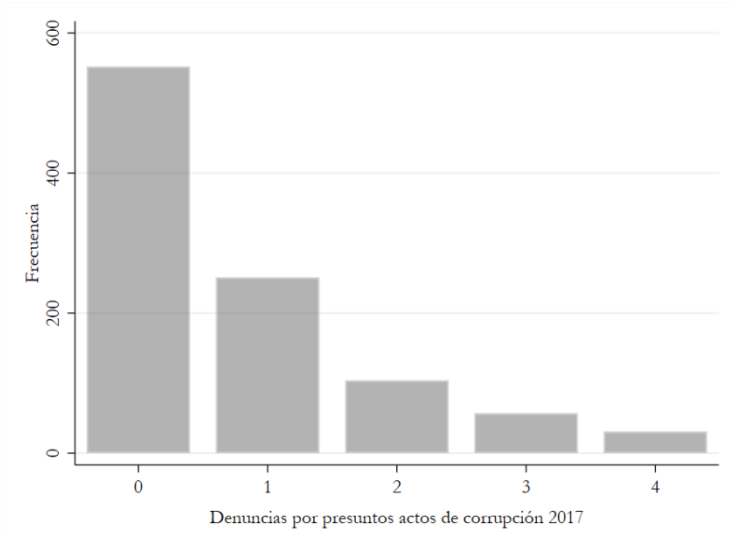
Test

Procesos





Denuncias



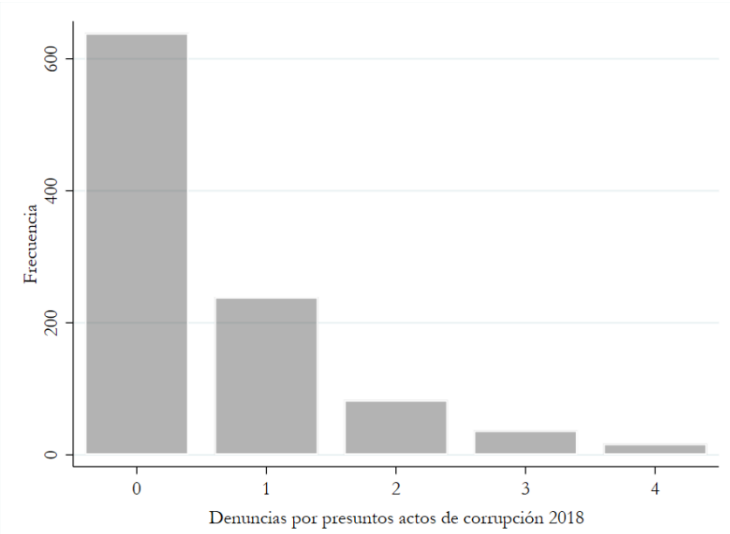


Tabla 2. Delitos contra la administración pública

Delito	Norma	2008 2011	2012 2015	2016 2019	
		casos	casos	casos	Total
Asociacion Para La Comision De Un Delito	Art. 434 C.P.	5	6	2	13
Cohecho Impropio	Art. 406 C.P.	14	19	6	44
Cohecho Por Dar U Ofrecer	Art. 407 C.P.	355	399	188	1032
Cohecho Propio	Art. 405 C.P.	54	67	30	171
Concusión	Art. 404 C.P.	178	153	60	475
Enriquecimiento Ilicito	Art. 412 C.P.	17	24	8	56
Interes Indebido En La Celebracion De Contratos	Art. 409 C.P.	40	25	7	91
Peculado Por Apropiacion	Art. 397 C.P.	427	245	41	921
Trafico De Influencias De Servidor Publico	Art. 411 C.P.	9	2	1	16
Total		1099	940	343	2819

Tabla 3. Estadísticas descriptivas variables objetivo

variable		Base	Obs	Mean	Sd	Min	Max
Procesos	Estimación Periodo electoral	Train	1,060	1.701	2.989	0.0	24.7
		Test	984	1.093	1.249	0.0	6.5
	Estimación anual	Train	995	1.266	1.915	0.0	10.0
		Test	1,015	1.177	1.806	0.0	9.0
Denuncias	Estimación Periodo electoral	Train	1,060	0.358	0.722	0.0	6.0
		Test	984	0.999	1.154	0.0	5.5
	Estimación anual	Train	995	0.758	1.054	0.0	4.0
		Test	1,015	0.575	0.910	0.0	4.0

Predictores

Predicción Anual: 102 variables – Predicción Periodo Electoral: 159 variables

Variables de Desempeño Económico	Variables de Desempeño Político	Variables de Gobierno	Variables de Características Generales
DATLAS - TERRIDATA	MOE – CEDE	TERRIDATA – CEDE	CEDE
+ Número de Establecimientos Industriales + Salario mensual promedio formal + Balanza comercial + Exportaciones per cápita +Porcentaje de la población ocupada formalmente	+ Partido político del alcalde +Permanencia partido en el poder +Indicador de fraude electoral + Indicador de atipicidad en votos nulos + Indicador de limitación a la competencia electoral	+ Desempeño integral del Municipio + Porcentaje de Ingresos corrientes +Categoría del municipio + Porcentaje de gasto destinado a inversión + Regalías per cápita + Indicador de Desempeño Fiscal +MDM	+ Indicador de exposición al conflicto armado +Indicador de calidad de la educación +población + Porcentaje ruralidad + Distancia a la capital departamental + Distancia a Bogotá + Altura + Densidad poblacional

Tabla 4. Correlación entre variables objetivos y algunas variables

	Procesos	Denuncias	Procesos t-1	población	Conflicto Index	Educación Index	Establecimientos
Procesos	1						
Denuncias	0.3769***	1					
Procesost_1	0.6332***	0.3797***	1				
población	0.5593***	0.3839***	0.5538***	1			
Conflicto_Ind	0.2118***	0.0363	0.2136***	0.3106***	1		
Educación_Ind	0.4148***	0.3748***	0.4115***	0.6940***	0.3134***	1	
Establecimientos	0.4951***	0.3863***	0.5093***	0.6391***	0.0901***	0.5798***	1

Metodología:

La investigación comprende las siguientes fases de trabajo y se realiza para la predicción por ciclo electoral y anual:

Preparación	Entrenamiento	Evaluación	Predicción	Riesgo
+Recolección de datos. + Selección de variables utilizando validación cruzada. +Tratamiento de missing values: Missing Forest. +Escalar las variables predictoras.	+Entrenar distintos algoritmos en varias combinaciones de hiperparámetros: Random Search. +Validar cuál es el error (RMSE) de cada combinación: validación cruzada.	+Evaluar el desempeño de la mejor combinación de hiperparámetros para cada uno de los algoritmos. +Seleccionar el mejor algoritmo de predicción.	+Re-entrenar con la muestra de evaluación y predecir las variables de interés en un periodo futuro.	+Comparar el resultado de la predicción con su histórico y demás variables contextuales para identificar municipios atípicos. +Agregar las variables en un indicador compuesto.

$$\tilde{Y}_{t,i} = f(X_{t-1,i}, X_{t-2,i}, \ldots, X_{t-k,i}; \tilde{Y}_{t-1,i}, \tilde{Y}_{t-2,i}, \ldots, \tilde{Y}_{t-k,i}, e_{t,i})$$

Algoritmos:

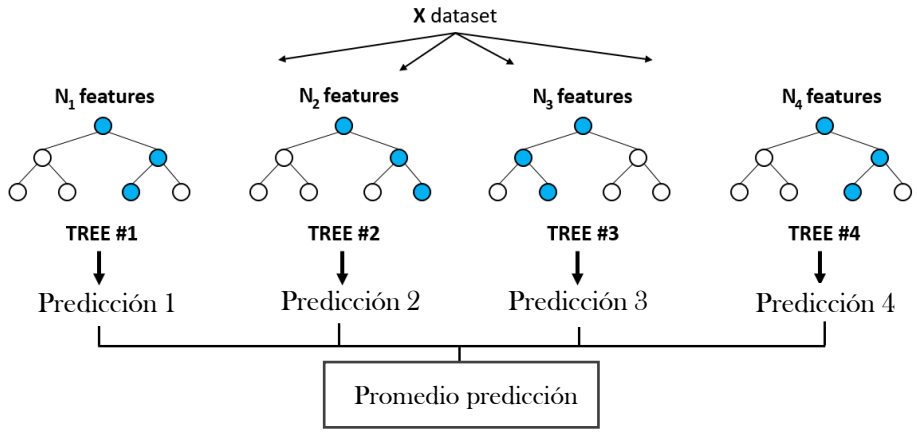
Gradient Boosting Machine

Árbol de decisión

Regresión regularizada

$$\hat{y}_i = \Phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

Random Forest



Train: 2017
Test: 2018

Tabla 3. Desempeño de los Algoritmos: Predicciones Anuales

			Muestra de Entrenamiento			Muestra de Evaluación		
Algoritmo	Variable de interés	Combinaciones evaluadas	RMSE Validación	SD	RMSE/SD	RMSE Evaluación	SD	RMSE/SD
XGBOOST árboles	Procesos	16215	**1.465**	1.915	**0.765**	1.578	1.806	0.874
	Denuncias	10000	0.970	1.054	0.920	**0.898**	0.910	**0.986**
XGBOOST regresión	Procesos	50000	1.482	1.915	0.774	**1.440**	1.806	**0.797**
	Denuncias	50000	**0.949**	1.054	**0.901**	0.916	0.910	1.006
Random Forest	Procesos	5000	1.512	1.915	0.789	1.445	1.806	0.800
	Denuncias	5000	0.969	1.054	0.919	0.919	0.910	1.010
Regresión Lineal	Procesos	1	1.565	1.915	0.817	3.587	1.806	1.986
	Denuncias	1	0.994	1.054	0.943	1.021	0.910	1.122

Train: 2012-2015
Test: 2016-2019

Tabla 2. Desempeño de los Algoritmos: Predicciones por Periodo Electoral

			Muestra de Entrenamiento			Muestra de Evaluación		
Algoritmo	Variable de interés	Combinaciones evaluadas	RMSE Validación	SD	RMSE/SD	RMSE Evaluación	SD	RMSE/SD
XGBOOST árboles	Procesos	10,000	1.847	2.985	0.619	2.282	1.249	1.827
	Denuncias	10,000	0.651	0.722	0.901	1.219	1.154	1.056
XGBOOST regresión	Procesos	25,000	1.867	2.985	0.626	**1.611**	1.249	**1.290**
	Denuncias	25,000	**0.647**	0.722	0.897	1.237	1.154	1.072
Random Forest	Procesos	5000	**1.843**	2.985	0.617	2.112	1.249	1.691
	Denuncias	5000	**0.647**	0.722	0.896	**1.0575**	1.154	**0.916**
Regresión Lineal	Procesos	1	3.29	2.985	1.102	2.125	1.249	1.702
	Denuncias	1	0.7675	0.722	1.063	1.1152	1.154	0.966

Gráfico 1. Importancia relativa de variables explicativas: Xgboost regresión variable Procesos

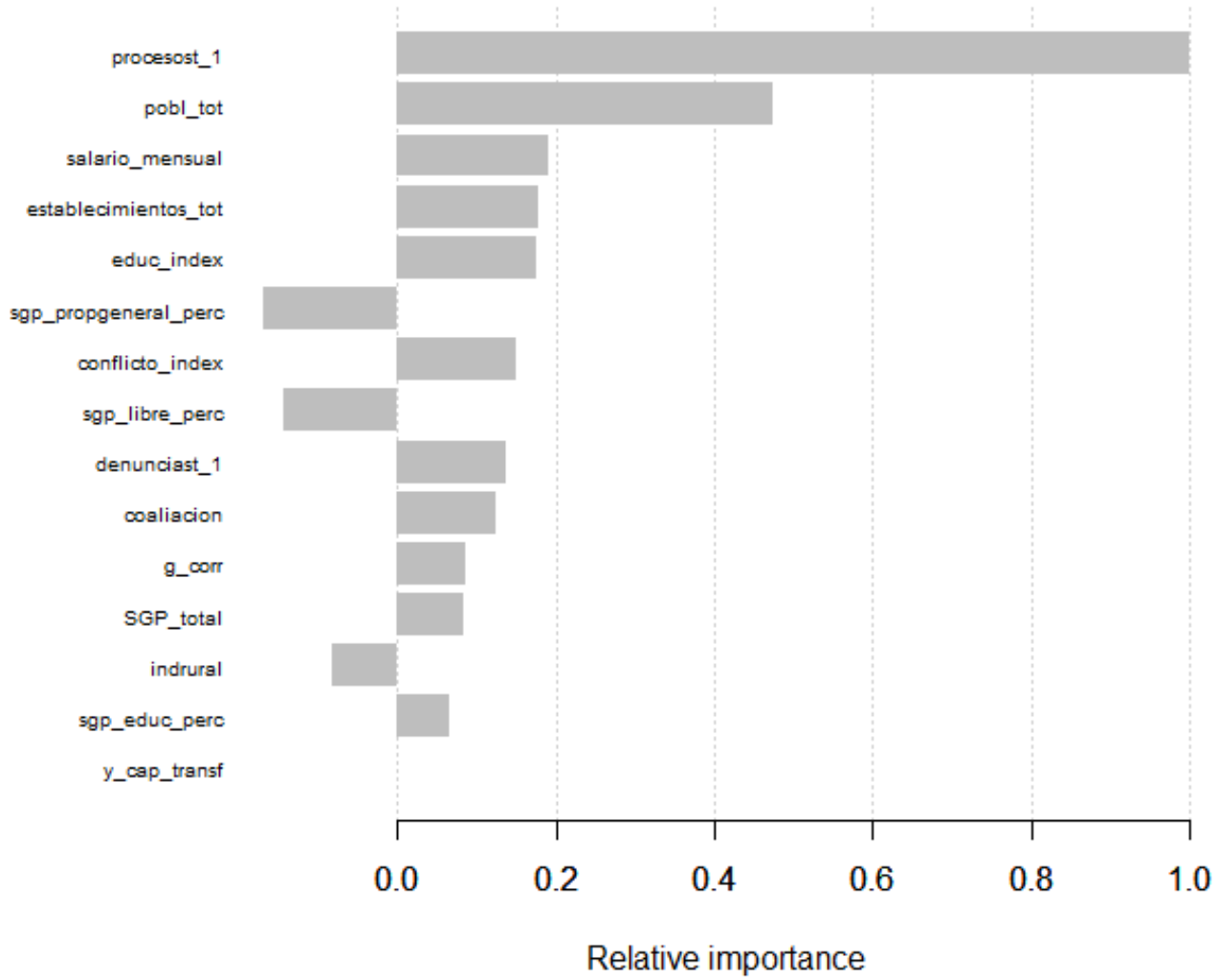
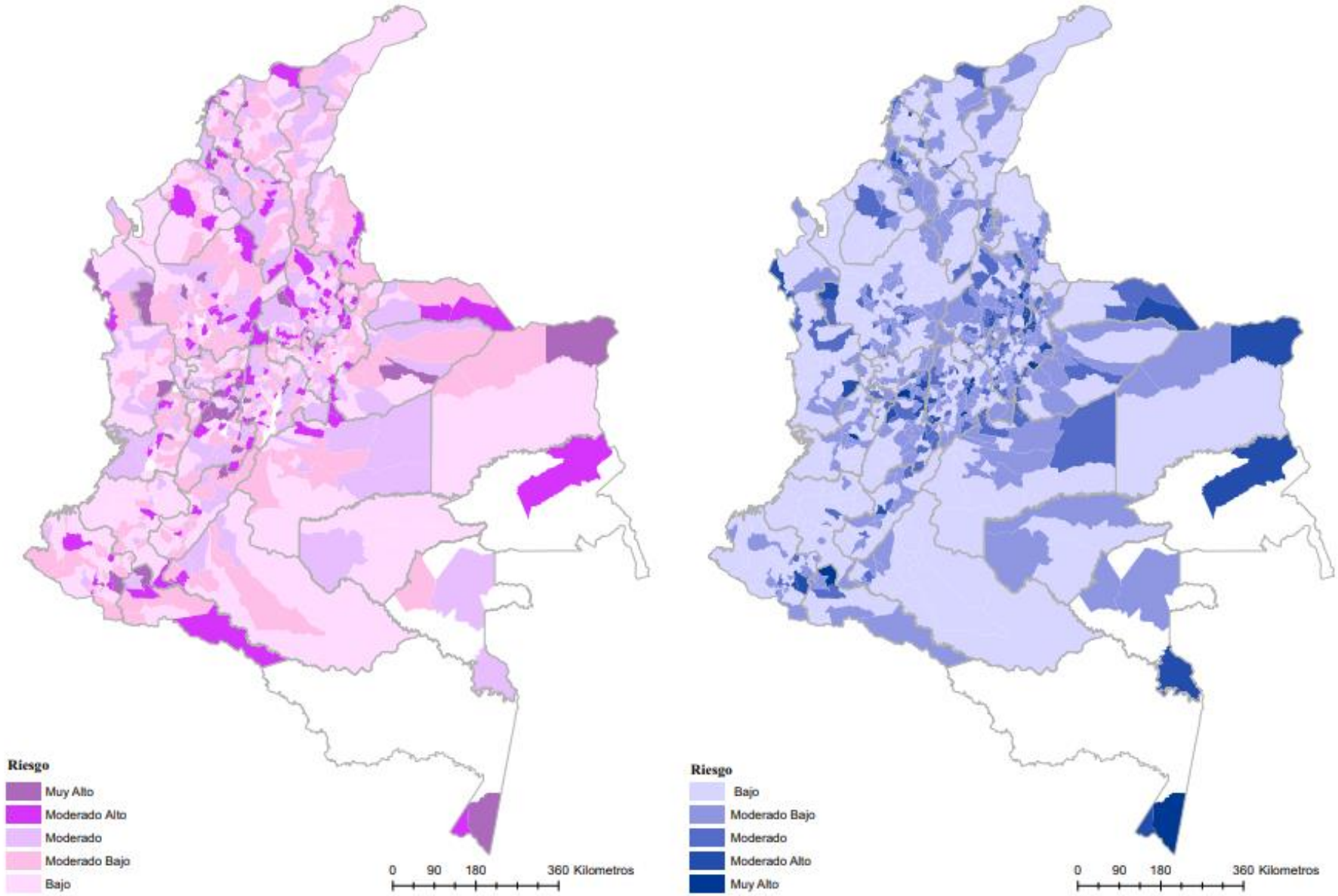


Gráfico 2. Clasificación de riesgo de corrupción por Clusters (izquierda) y clasificación de riesgo de corrupción por Componentes Principales



Gracias por su atención