

Estimación de población étnica a niveles subnacionales

Aprovechamiento de información auxiliar



**El futuro
es de todos**

**Gobierno
de Colombia**



Fases para la estimación de la población étnica

Fase I

Determinar la participación de personas autorreconocidas como pertenecientes a un grupo étnico a nivel nacional.

La estimación de la ECV refleja de forma idónea el comportamiento de las personas que se autorreconocen en las categorías étnicas en el CNPV 2018.

Fase II

Asignar pertenencia étnica a las personas: Dentro del total de personas efectivamente censadas en el CNPV 2018.

Fase III

Al contar con la georreferenciación de todas las personas efectivamente censadas, es posible **determinar la participación de cada grupo étnico a niveles de desagregación inferiores al nivel municipal** (manzanas o sectores rurales). Esta información permite aplicar la metodología propuesta por (Stevens, 2015): aplicar un modelo de aprendizaje (RandomForest Regression o Gradient-Boosted Trees Regression).

Fase IV

Distribuir proporcionalmente la población omitida no asignada en Fase 2 y 3 por pertenencia étnica: **Tabla Cuadrada**



Fase 1

Los resultados de la encuesta de calidad de vida 2018 (ECV) para población étnica a nivel nacional en el año 2018 cuentan con un coeficiente de variación que permite realizar inferencias sobre la participación poblacional.

Grupo Étnico	Partencia Étnica	Participación	cv
1	indígena	4,34%	3,2
2	Gitano(a) o Rrom	0,01%	38,7
3	Raizal del archipiélago de San Andres, Providencia y Santa Catalina	0,04%	12,5
4	Palenquero(a) de San Basilio	0,02%	25,2
5	Negro(a), mulato(a), afrodescendiente, afrocolombiano(a)	9,28%	2,5
6	Ningún grupo étnico	86,31%	0,3

Asignación pertenecía étnica usando modelos predictivos de clasificación

Fase II



**El futuro
es de todos**

**Gobierno
de Colombia**



Existen 595.586 personas que no cuentan con información de autorreconocimiento étnico. La propuesta consiste en adaptar la metodología presentada en (Voicu, 2018), que mediante modelos de machine learning bayesianos determinan la pertenencia étnica de las personas usando los nombres como covariados.

Grupo Étnico	Partencia Étnica	Total	Participación
1	Indígena	1.905.617	4,3%
2	Gitano(a) o Rrom	2.649	0,0%
3	Raizal del archipiélago de San Andres, Providencia y Santa Catalina	25.515	0,1%
4	Palenquero(a) de San Basilio	6.637	0,0%
5	Negro(a), mulato(a), afrodescendiente, afrocolombiano(a)	2.950.072	6,7%
6	Ningún grupo étnico	38.678.341	87,6%
9	Sin información	595.586	1,3%
Total		44.164.417	100,0%

Cálculo de probabilidades

Esta asignación se da de manera probabilística usando las siguientes probabilidades condicionales:

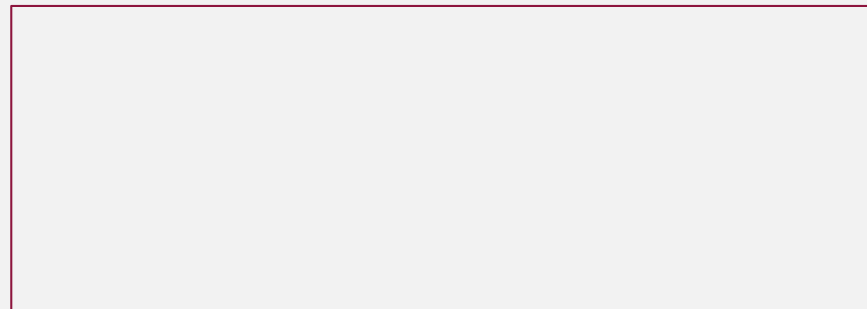
- , probabilidad de pertenecer a un grupo étnico, dado el apellido de la persona, con el grupo étnico y el apellido de la persona.
- , probabilidad de llamarse de determinada manera dado que se pertenece a un grupo étnico, con el nombre de la persona y el grupo étnico
- , probabilidad de residir en determinada ubicación geográfica dado que se pertenece a un grupo étnico, con desagregación geográfica y el grupo étnico



Apellido	Grupo Étnico	P(r a)
AGUILAR	1	0.0062
AGUILAR	4	0.0012
AGUILAR	5	0.9589
AGUILAR	6	0.0337
MOSQUERA	1	0.0017
MOSQUERA	3	0.0001
MOSQUERA	4	0.0002
MOSQUERA	5	0.9879
MOSQUERA	6	0.0100
LOZANO	1	0.0036
LOZANO	5	0.9749
LOZANO	6	0.0215

Grupo Étnico	Nombre	P(r n)
	5 BAYRON	0.9000
	6 BAYRON	0.1000
	1 DILMAR	0.4000
	5 DILMAR	0.6000
	1 JUANA	0.0252
	5 JUANA	0.9627
	6 JUANA	0.0121

Grupo Étnico	Municipio	P(t r)
	1	0.0402
	2	0.0000
	3	0.0004
	4	0.0002
	5	0.9275
	6	0.0318



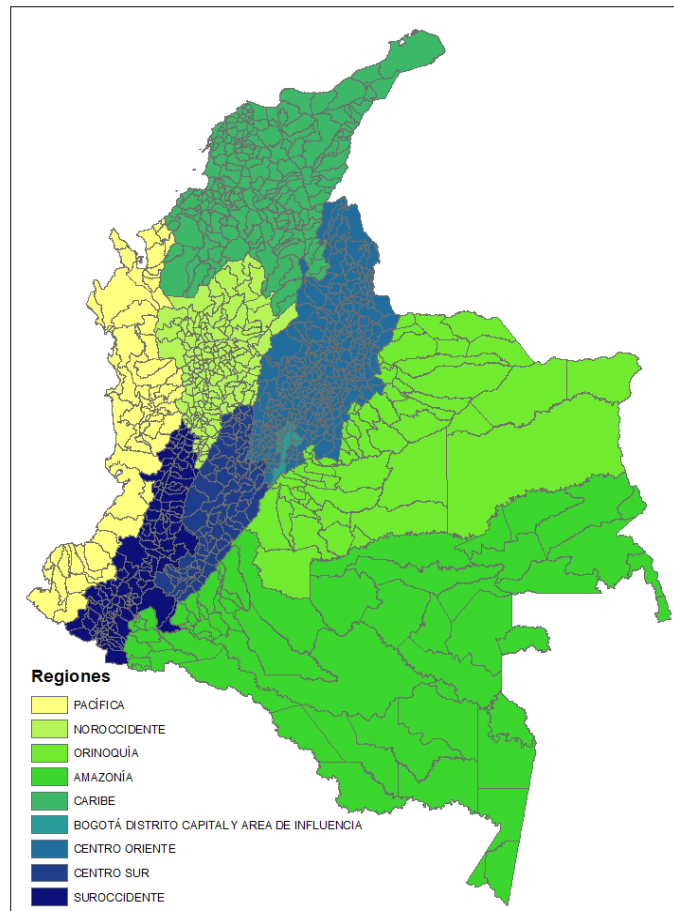


Modelo de Random Forest diagnóstico

Se tiene como insumo un conjunto de datos que cuenta con **43.534.030** filas y 24 variables que representan las probabilidades condicionales observadas.

Luego se ajusta un modelo de aprendizaje de máquinas para clasificar estas personas en una de las 3 categorías objeto del análisis.

Por ultimo, se realiza una partición del insumo en un conjunto de aprendizaje y un conjunto de prueba, con una distribución 80-20, y se ajusta el modelos para cada una de las 9 regiones construidas.

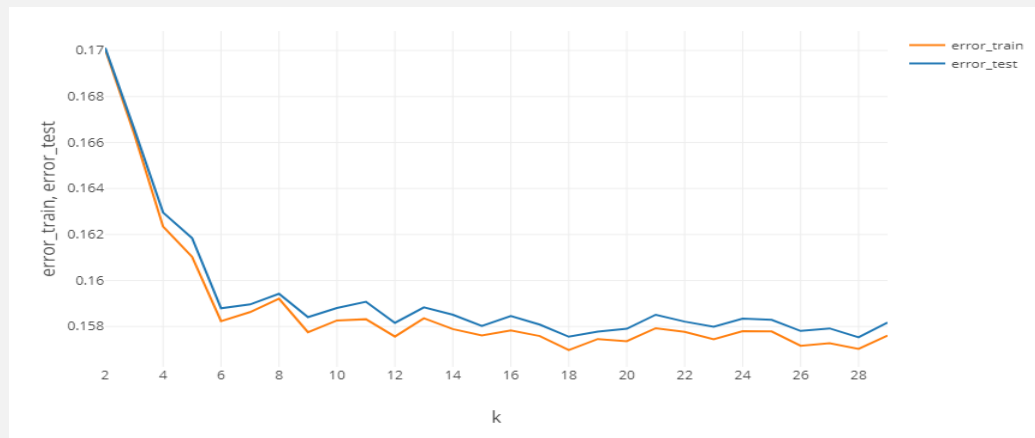




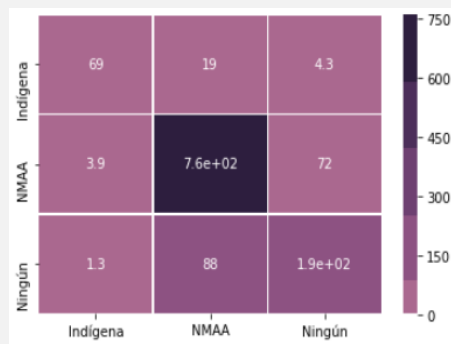
Modelo Región Pacífica

El ajuste en la región Pacífica es el que presenta mejor comportamiento, como se puede observar en sus matrices de confusión, alcanzando un precisión del **84%** cuando el parámetro k es 18.

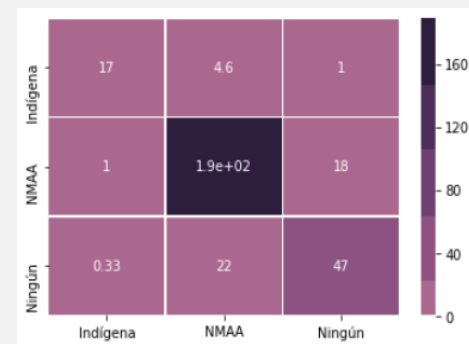
Fuente: DANE Dirección de censos y demografía CNPV 2018



Entrenamiento



Prueba

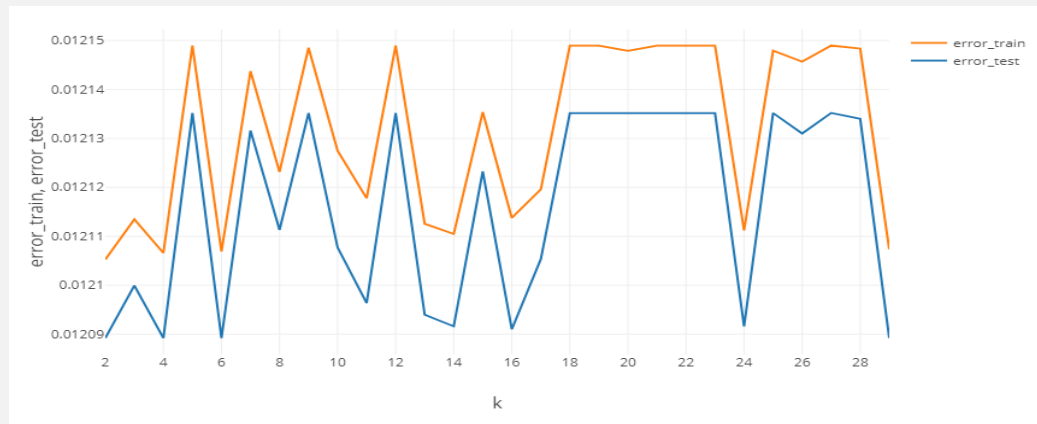


Modelo Región Bogotá y área de influencia

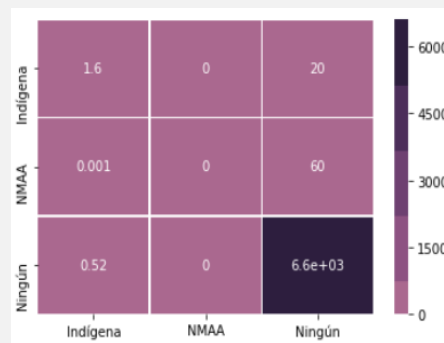
La región no alcanzan un poder predictivo aceptable, puesto que, aún con una precisión del **99%**, la columna de ningún grupo étnico en las matrices de confusión acumula el mayor porcentaje de personas.

No es adecuado usar la predicción del modelo para esta región.

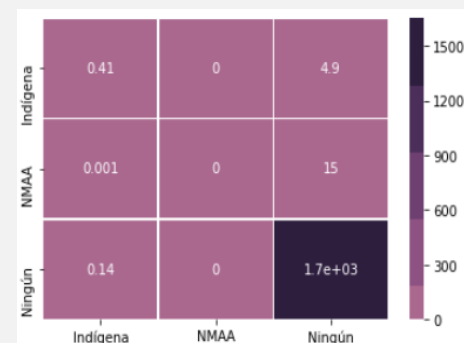
Fuente: DANE, Dirección de censos y demografía CNPV 2018.



Entrenamiento



Prueba





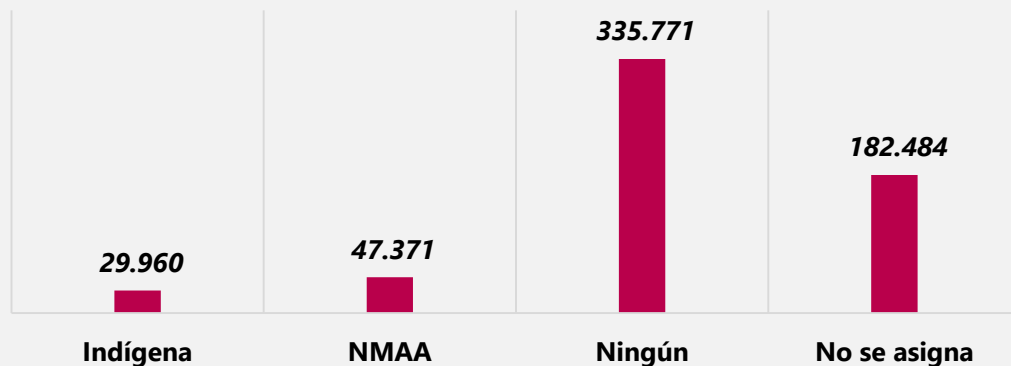
Los modelos presentan un desempeño aceptable y se recomienda su implementación para las regiones:

- Pacífica
- Noroccidente
- Orinoquía
- Amazonía
- Caribe
- Centro oriente
- Suroccidente

No se recomienda el uso para:

- Bogotá y área de influencia
- Centro sur

Predicción Fase II



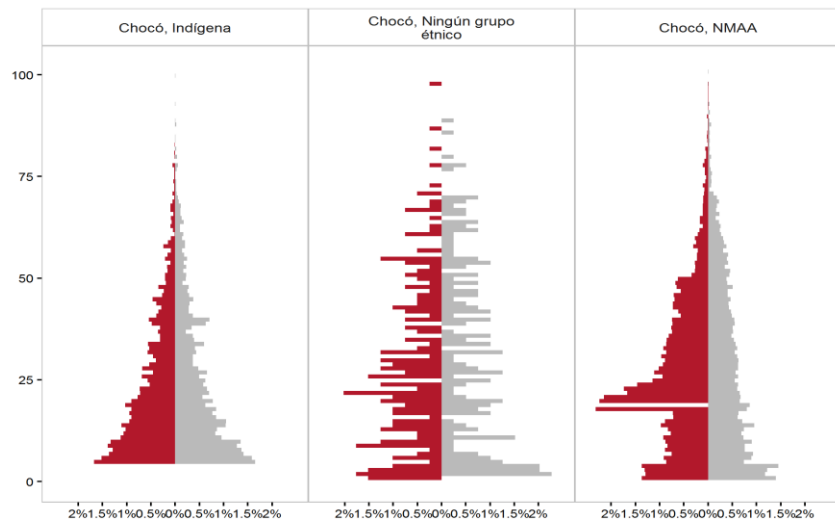
Municipios con mayor número de personas asignadas como indígenas y NMAA

Departamento	Municipio	Nmaa
Nariño	Tumaco	6,000
Choco	Riosucio	4,763
Valle del cauca	Buenaventura	4,169
Nariño	Barbacoas	2,634
Choco	Quibdo	2,105

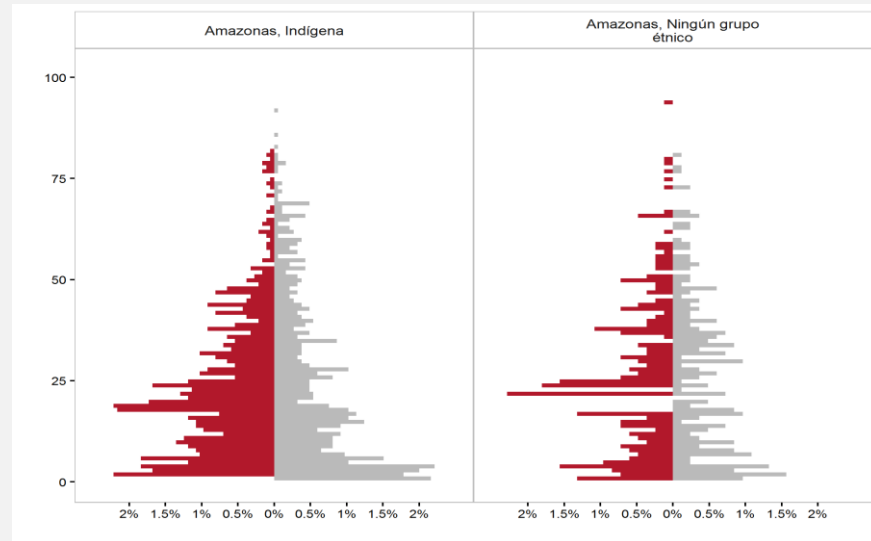
Departamento	Municipio	Indígena
Vichada	Cumaribo	2,060
Choco	Alto baudo	2,003
La guajira	Dibulla	1,940
Choco	Bojaya	1,397
Choco	Lloro	1,172



Código	Departamento	Etnia	2018
27	Chocó	Indígena	9,008
27	Chocó	NMAA	20,396
27	Chocó	Ningún	396



Código	Departamento	Etnia	2018
91	Amazonas	Indígena	1,851
91	Amazonas	Ningún	831





Fase III

Esta fase esta compuesta por 2 sub-procesos:

1.

Estimación de la omisión indígena del CNPV 2018 por fragmento de resguardo.

2.

Estimación de la omisión NMAA del CNPV 2018 por manzana y sección rural del MGN.

Estimación de la omisión del CNPV 2018 por fragmento de resguardo



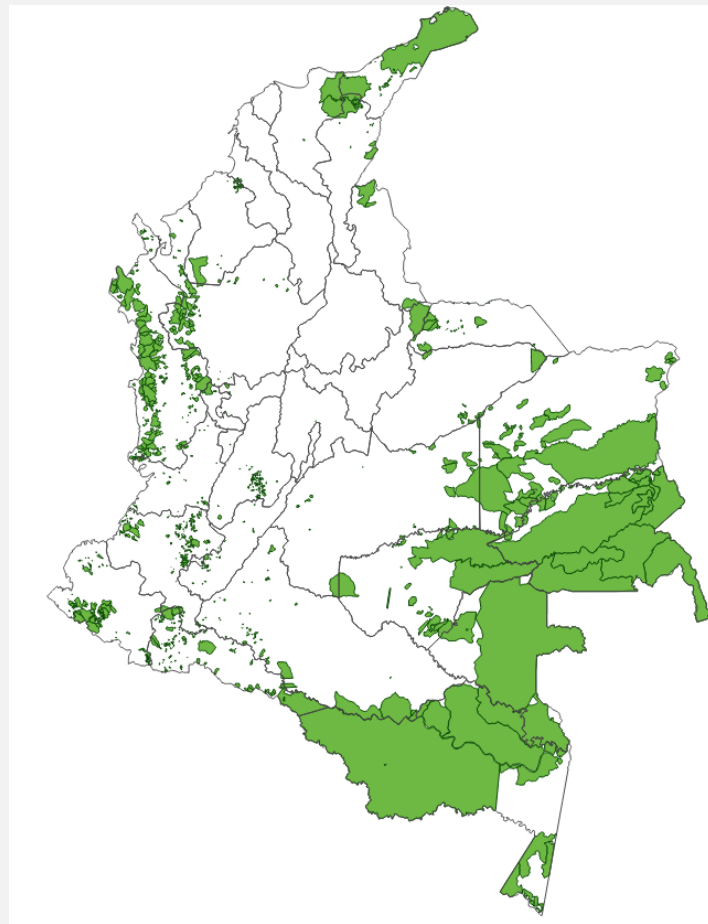
El futuro
es de todos

Gobierno
de Colombia



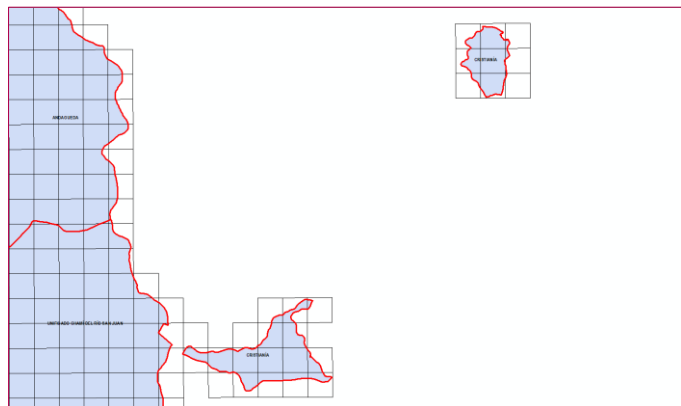
Los **813 resguardos** se distribuyen en **29** departamentos y **279** municipios.

Según resultados del CNPV 2018 con cálculo de omisión, encuentra población indígena en **28 departamentos, y 266 municipios**

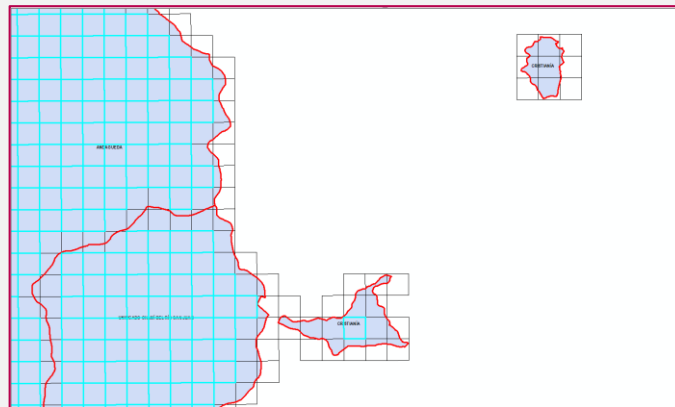




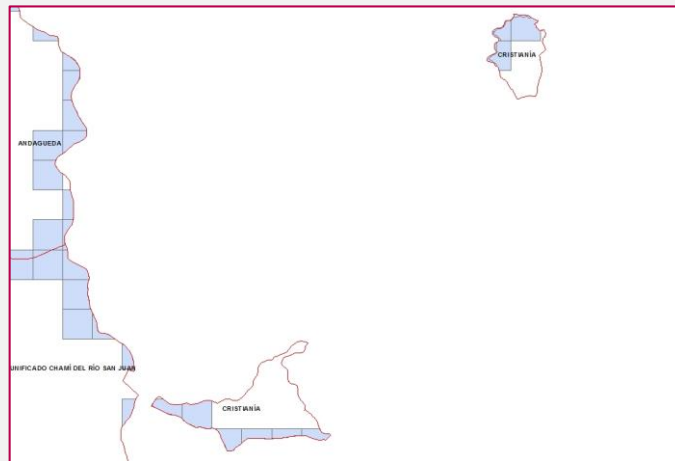
Grilla Circunscrita



Grilla Inscrita

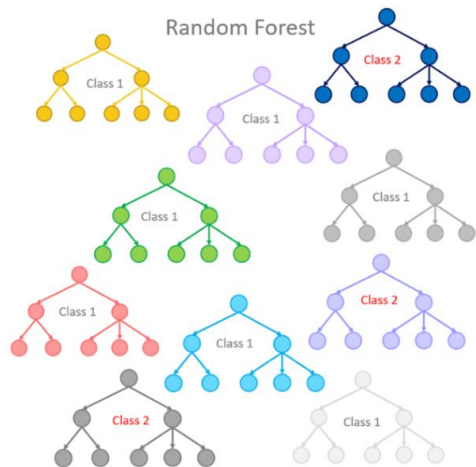


Fragmento de Grilla



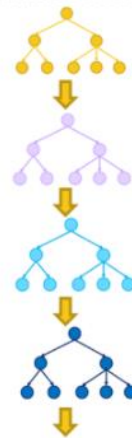


El **Random Forest** es conjuntos de árboles de decisión (Es una representación en forma de árbol cuyas ramas se bifurcan en función de los valores tomados por las variables y que terminan en una acción concreta); estos combinan muchos árboles de decisión para reducir el riesgo de sobreajuste.



El **Gradient-Boosted Trees** (GBT) son conjuntos de árboles de decisión. Los GBT entrenan iterativamente los árboles de decisión para minimizar una función de pérdida.

Gradient Boosted Trees



- Las dos metodologías permiten clasificación binaria, multiclase y regresión, utilizando características continuas y categóricas.



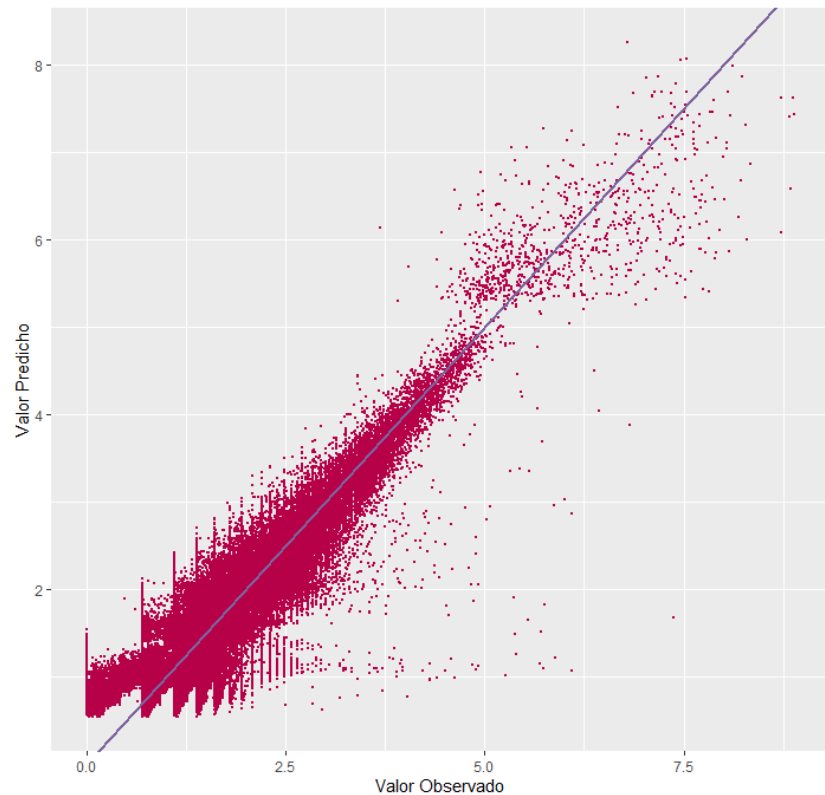
- **Variable dependiente:** logaritmo de las personas observadas en cada unidad de cobertura.
- **Variables independientes:** densidad de colegios (DEN_COLEG), distancia a centros poblados (DIST_CP), elevación (ELEVACION), luces nocturnas (LUCES_NOCT), índice de vegetación normalizado(NDVI), pendiente (PENDIENTE), densidad de construcciones (DEN_CONSTR), densidad de IPS (DEN_IPS), porcentaje de agua superficial (PORCENT_AG), total unidades de vivienda (TOTAL_U_VI), total de personas en VOCPA (A), total de personas estimados por subenumeración (S), total de personas estimada en zonas no visitadas o incompletas (Z) , total de hogares (hogares).



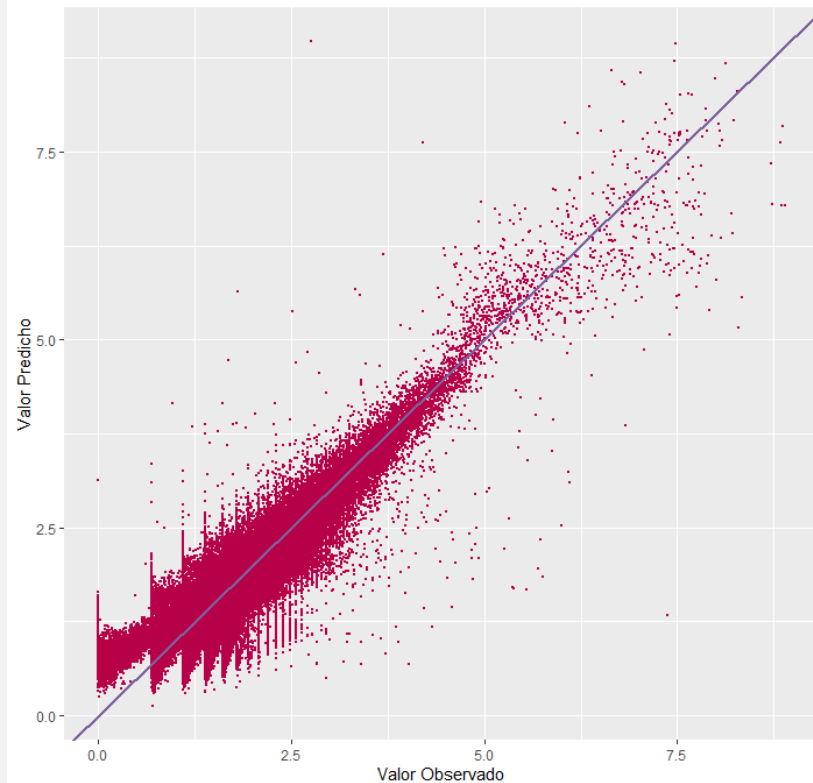
- **Conjunto de aprendizaje:**
608.371 unidades de cobertura en rural disperso
- **Conjunto de Prueba:**
261.634 unidades de cobertura en rural disperso
- **Implementación:**
Datbricks con Pyspark

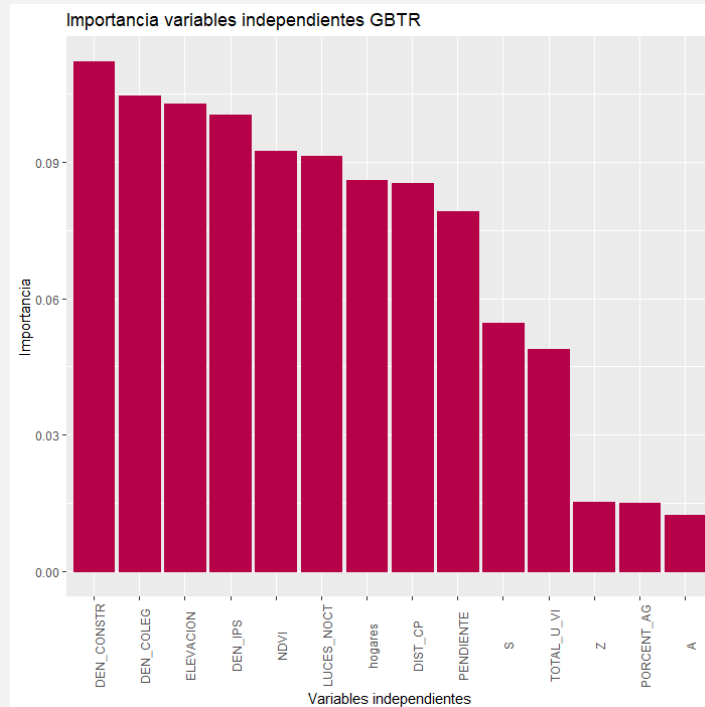
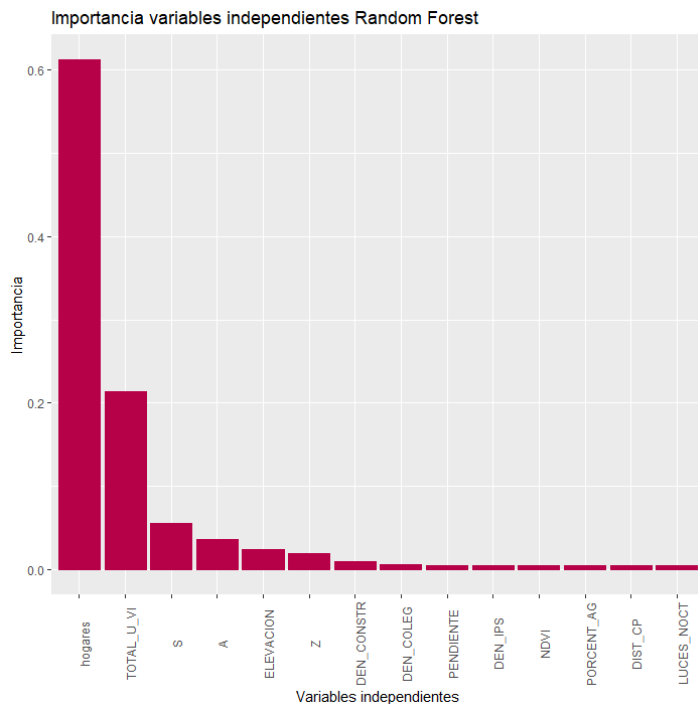


Predicción Random Forest 45 arboles 15 ramas



GBTR 40 iteraciones 10 ramas

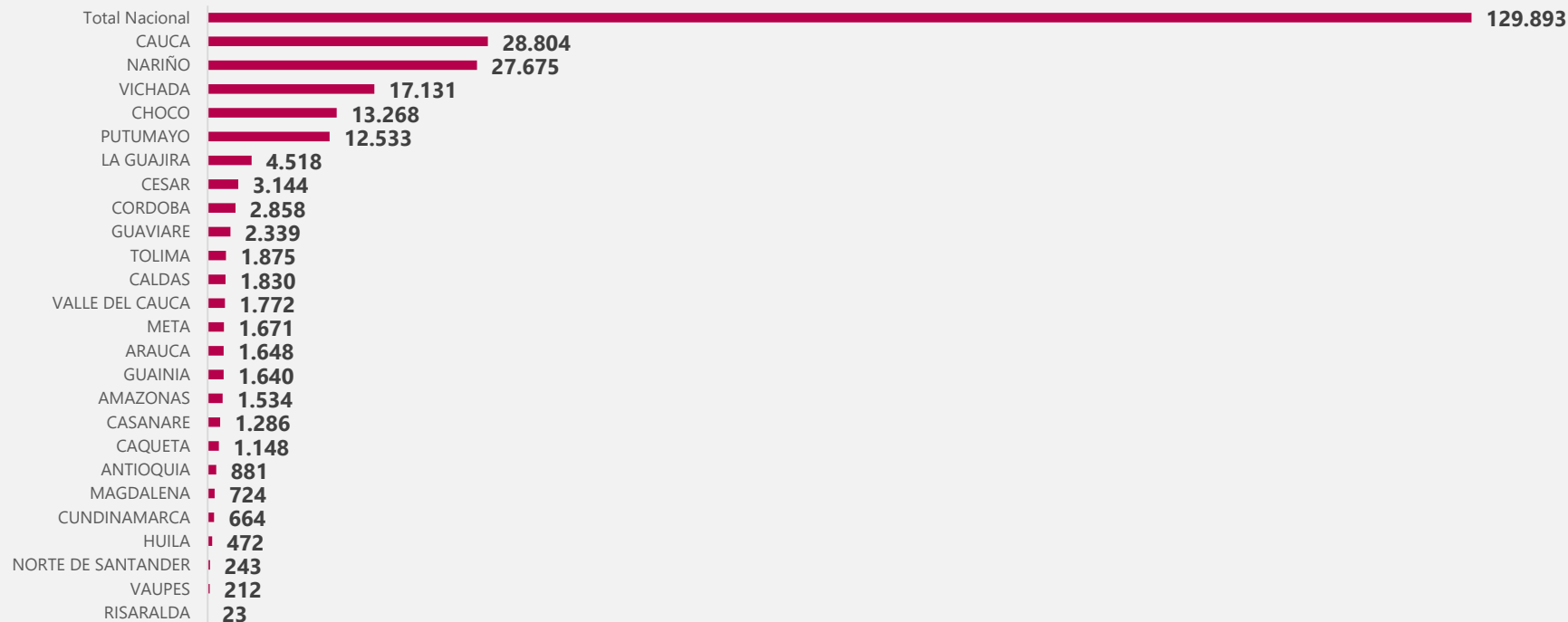




Diagnóstico		
	Random forest	GBTR
RMSE	0.457891	0.453902
R^2	0.704725	0.709847



Total de personas omitidas indígenas en resguardos a nivel departamental



***Estimación de la participación
Negra, Mulata, afrodescendiente
y afrocolombiana usando
aprendizaje de maquinas***





Estimación de la participación Negra, Mulata, afrodescendiente y afrocolombiana usando aprendizaje de maquinas

Objetivo fase III

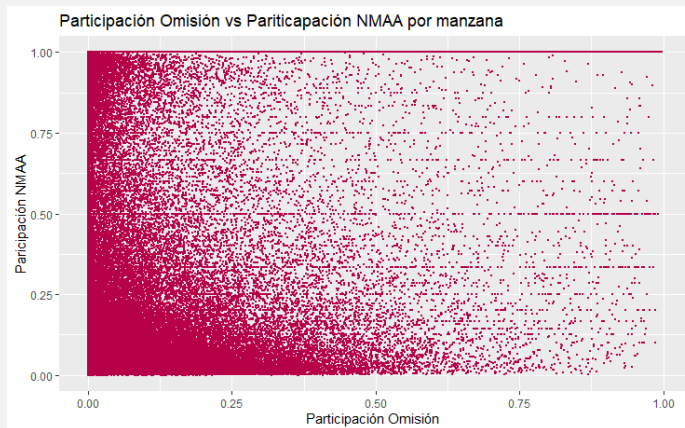
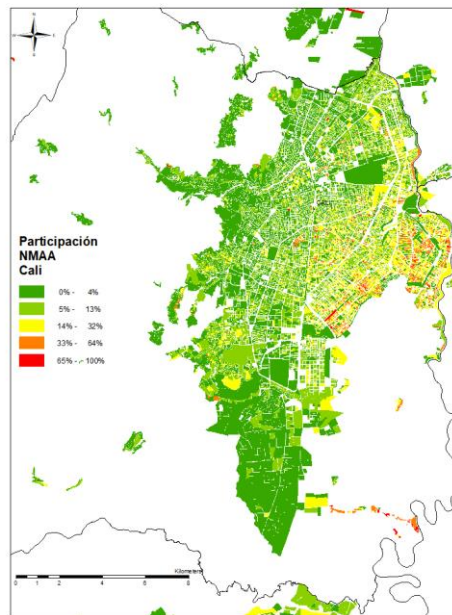
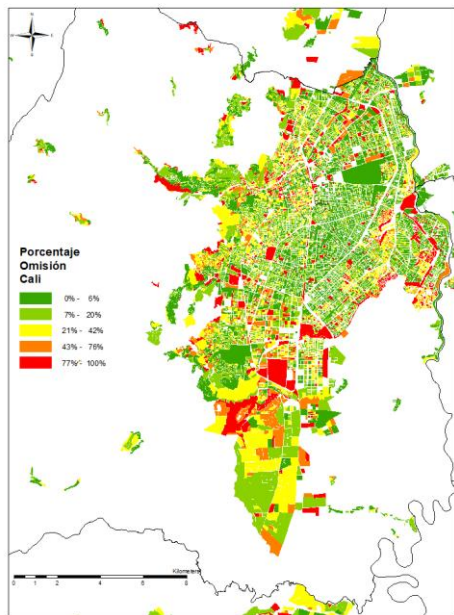
Determinar la participación de la población negra, mulata, afrocolombiana o afrodescendiente dentro del 8.5% de personas omitidas en el CNPV 2018.

Objetivos específicos

1. Determinar el porcentaje de omisión y las participaciones censales de la población NMAA a nivel de manzana y sector rural.
2. Construir un conjunto de aprendizaje a partir de los indicadores obtenidos anteriormente.
3. Entrenar un modelo predictivo que determine la participación NMAA en las manzanas o sectores.
4. Diagnosticar los resultados del modelo empleado.
5. Predecir sobre las manzanas o sectores rurales con porcentaje de omisión mayor a 0 y excluyendo al conjunto de aprendizaje.
6. Agregar las predicciones a nivel municipal y analizar los resultados.



1. Determinar el porcentaje de omisión y las participaciones censales de la población NMAA a nivel de manzana o sector rural



Cada punto representa una manzana o sector rural en las que por lo menos una persona se autorreconocen como NMAA. Se tienen 162.556 puntos.

*Participación NMAA con resultados de Fasell



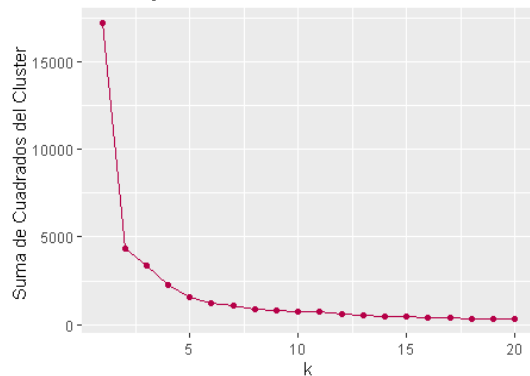
Modelo empleado

1. Realizar grupos de regiones, con el fin de no distribuir el comportamiento de zonas con altas participaciones en zonas que temáticamente no cuentan con estos grupos.
2. Convertir los covariados en indicadores a nivel manzana, se cuentan con 62 indicadores como tamaño promedio del hogar, proporción, de alfabetismos, proporción de viviendas, por tipo de piso, etc.
3. Determinar un conjunto de aprendizaje, de prueba y predicción para cada uno de los anteriores grupos.
4. Entrenar tres modelos para cada uno de los grupos y predecir.
5. Ajustar la omisión de manzana armonizándola con la desagregación cabecera - resto.

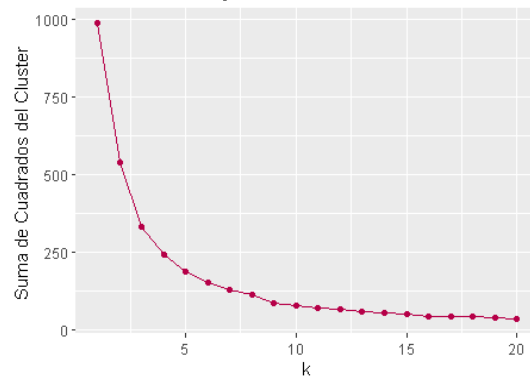




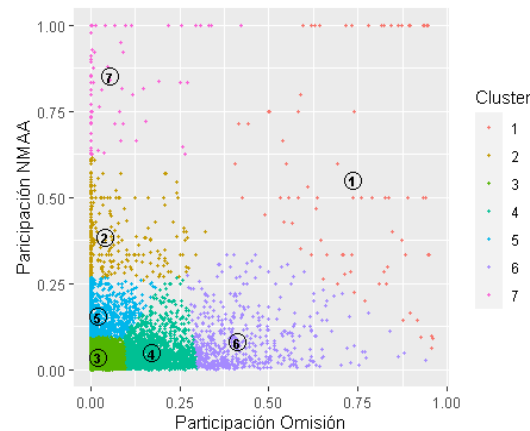
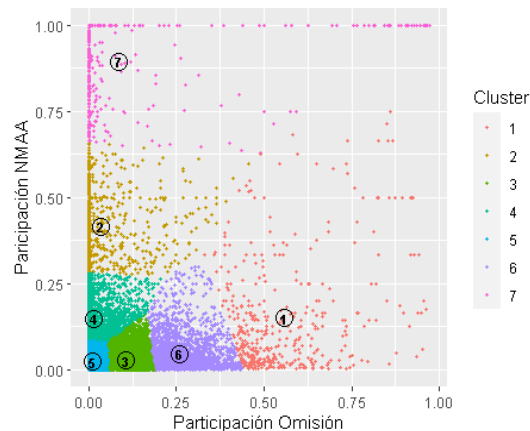
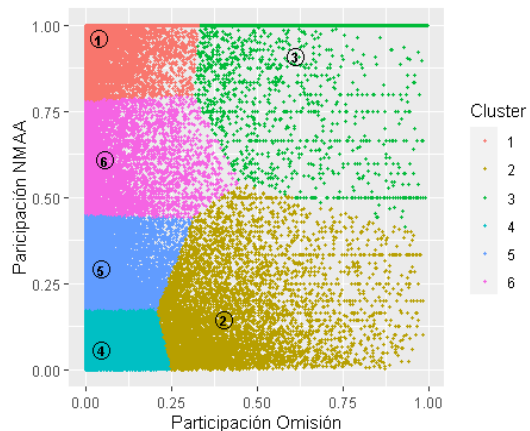
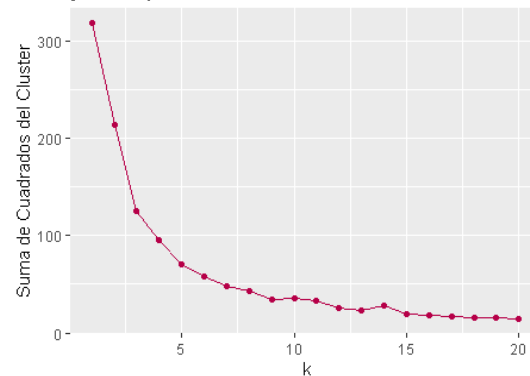
Pacífico,
Caribe y Suroccidente



Noroccidente, Bogotá,
Centro Oriente y Centro Sur



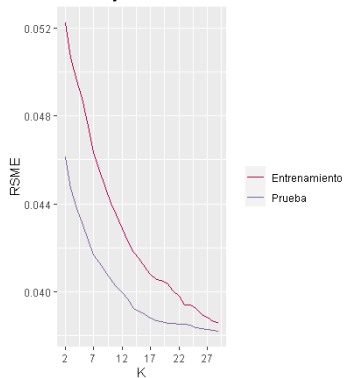
Amazonas
y Orinoquia



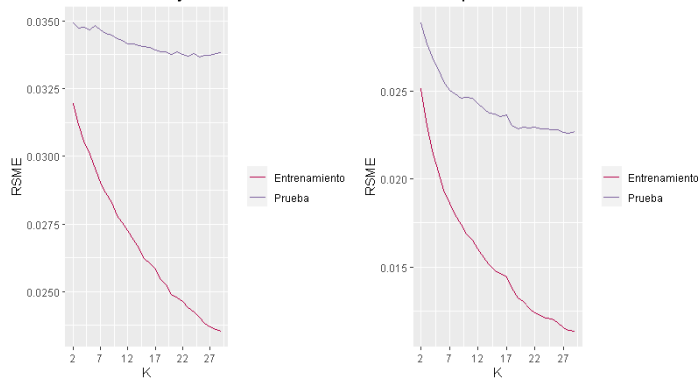


INFORMACIÓN PARA TODOS

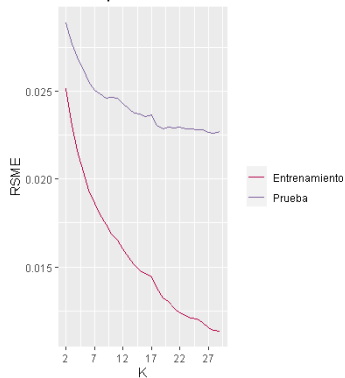
Pacífico,
Caribe y Suroccidente



Noroccidente, Bogotá,
Centro Oriente y Centro Sur

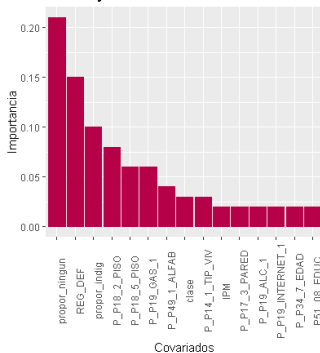


Amazonas y
Orinoquía

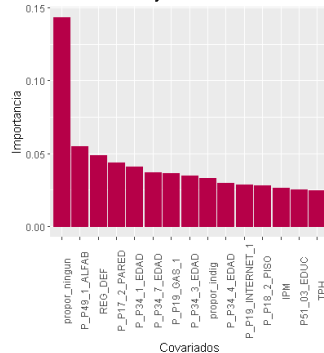


Región	RMSE	R ²
Pacífica, Atlántica y Suroccidente	0,043	0,9879
Noroccidente, Bogotá y área de influencia, Centro oriente, Centro sur	0,0365	0,8942
Amazonía y Orinoquía	0,0237	0,9217

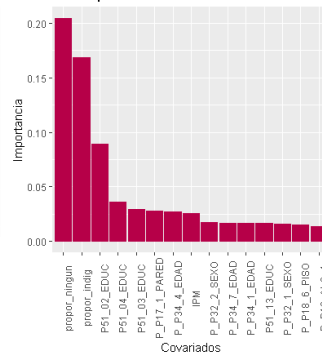
Pacífico,
Caribe y Suroccidente



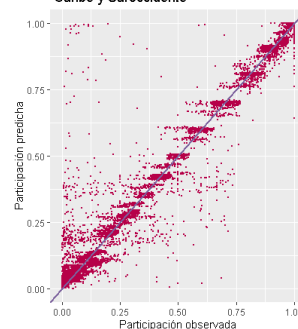
Noroccidente, Bogotá,
Centro Oriente y Centro Sur



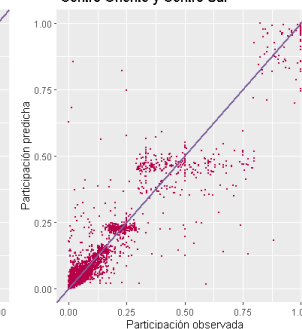
Amazonas y
Orinoquía



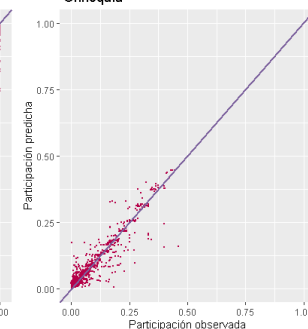
Pacífico,
Caribe y Suroccidente



Noroccidente, Bogotá,
Centro Oriente y Centro Sur



Amazonas y
Orinoquía





Predicciones y resultados

1. Los resultados del modelo son condicionados a las participaciones observadas, es decir si el modelo predice una menor proporción de NMAA dentro de la manzana se toma la proporción observada.
2. Dado que el conjunto de aprendizaje tiene niveles bajos de omisión (**147.888** manzanas o sectores rurales), se considera que el comportamiento étnico dentro de las personas omitidas debe conservar la estructura observada de las personas efectivamente censadas, por lo tanto, se aplica la participación étnica observada a la población omitida.
3. Los modelos predicen la participación dentro de las manzana (**202.598** con por lo menos una persona omitida), por lo tanto el ajuste al volumen se realiza como:

Donde:

: es la proporción de NMAA observada en la manzana

: es la predicción de la proporción de NMAA en manzana.

Número de personas

4. Por ultimo, en caso en que , el número de personas NMAA omitidas en a manzana es igual al número de personas omitidas.



INFORMACIÓN PARA TODOS

DPTO	Departamento	Población Total	Población Efectivamente censada	Población Omitida CNPV 2018	NMAA Declaración CNPV 2018	Estimación Total NMAA Fase II	Estimación Total NMAA Omitido Fase III	Población NMAA ajustada	Participación NMAA CNPV 2018	Participación NMAA ajustada
05	Antioquia	6,407,102	5,974,788	432,314	311,289	1,390	35,416	348,095	5.2%	5.4%
08	Atlántico	2,535,517	2,342,265	193,252	138,812	329	24,141	163,282	5.9%	6.4%
11	Bogotá, D.C.	7,412,566	7,181,469	231,097	65,656	-	31,885	97,541	0.9%	1.3%
13	Bolívar	2,070,110	1,909,460	160,650	314,835	587	26,867	342,289	16.5%	16.5%
15	Boyacá	1,217,376	1,135,698	81,678	4,169	-	5,007	9,176	0.4%	0.8%
17	Caldas	998,254	923,472	74,783	14,579	1	7,403	21,983	1.6%	2.2%
18	Caquetá	401,849	359,602	42,247	5,043	-	69	5,112	1.4%	1.3%
19	Cauca	1,464,488	1,243,503	220,985	245,183	3,745	38,367	287,295	19.7%	19.6%
20	Cesar	1,200,574	1,098,577	101,997	142,233	284	14,360	156,877	12.9%	13.1%
23	Córdoba	1,784,783	1,555,596	229,187	102,251	534	36,758	139,543	6.6%	7.8%
25	Cundinamarca	2,919,060	2,792,877	126,183	12,884	-	9,474	22,358	0.5%	0.8%
27	Chocó	534,826	457,412	77,414	337,442	20,396	61,132	418,970	73.8%	78.3%
41	Huila	1,100,386	1,009,548	90,838	5,027	-	7,428	12,455	0.5%	1.1%
44	La Guajira	880,560	825,364	55,196	60,256	20	15,774	76,050	7.3%	8.6%
47	Magdalena	1,341,746	1,263,788	77,958	106,143	12	13,856	120,011	8.4%	8.9%
50	Meta	1,039,722	919,129	120,593	8,710	-	157	8,867	0.9%	0.9%
52	Nariño	1,630,592	1,335,521	295,071	232,847	14,353	121,992	369,192	17.4%	22.6%
54	Norte de Santander	1,491,689	1,346,806	144,883	5,410	-	9,856	15,266	0.4%	1.0%
63	Quindío	539,904	509,640	30,264	6,036	-	3,264	9,300	1.2%	1.7%
66	Risaralda	943,401	839,597	103,804	16,608	13	7,992	24,613	2.0%	2.6%
68	Santander	2,184,837	2,008,841	175,996	22,554	-	18,046	40,600	1.1%	1.9%
70	Sucre	904,863	864,036	40,827	102,655	349	7,537	110,541	11.9%	12.2%
73	Tolima	1,330,188	1,228,763	101,424	5,110	-	9,035	14,145	0.4%	1.1%
76	Valle del Cauca	4,475,886	3,789,874	686,012	646,762	5,202	90,915	742,879	17.1%	16.6%
81	Arauca	262,174	239,503	22,671	9,994	-	292	10,286	4.2%	3.9%
85	Casanare	420,504	379,892	40,612	6,089	-	127	6,216	1.6%	1.5%
86	Putumayo	348,182	283,197	64,985	10,220	92	270	10,582	3.6%	3.0%
88	Archiipiélago de San Andrés	61,280	48,299	12,981	6,531	5	4,230	10,766	13.5%	17.6%
91	Amazonas	76,589	66,056	10,533	473	-	3	476	0.7%	0.6%
94	Guainía	48,114	44,431	3,683	451	-	8	459	1.0%	1.0%
95	Guaviare	82,767	73,081	9,686	2,980	59	97	3,136	4.1%	3.8%
97	Vaupés	40,409	37,690	3,107	270	-	1	271	0.7%	0.7%
99	Vichada	107,808	76,642	31,166	570	-	7	577	0.7%	0.5%
Total general		48,258,106	44,164,417	4,094,077	2,950,072	47,371	601,764	3,599,207	6.7%	7.5%

***Comparativo
departamental entre
Censo General 2005,
CNPV 2018 y ajuste al
CNPV 2018***



El futuro
es de todos

Gobierno
de Colombia



INFORMACIÓN PARA TODOS

DPTO	Departamento	Censo General 2005			CNPV 2018			Ajuste CNPV 2018		
		Población Efectivamente censada	NMAA Declaración 2005	Participación NMAA	Población Efectivamente censada	NMAA Declaración CNPV 2018	Participación NMAA	Población Total	Poblacion NMAA ajustada	Participación NMAA
5	Antioquia	5,601,507	593,174	10.6%	5,974,788	311,289	5.2%	6,407,102	348,095	5.4%
8	Atlántico	2,112,001	224,109	10.6%	2,342,265	138,812	5.9%	2,535,517	163,282	6.5%
11	Bogotá, D.C.	6,778,691	96,523	1.4%	7,181,469	65,656	0.9%	7,412,566	97,541	1.3%
13	Bolívar	1,836,640	491,364	26.8%	1,909,460	314,835	16.5%	2,070,110	342,289	16.7%
15	Boyacá	1,210,982	16,602	1.4%	1,135,698	4,169	0.4%	1,217,376	9,176	0.7%
17	Caldas	898,490	22,631	2.5%	923,472	14,579	1.6%	998,255	21,983	2.2%
18	Caquetá	337,932	11,661	3.5%	359,602	5,043	1.4%	401,849	5,112	3.0%
19	Cauca	1,182,022	255,839	21.6%	1,243,503	245,183	19.7%	1,464,488	287,295	20.4%
20	Cesar	878,437	105,273	12.0%	1,098,577	142,233	12.9%	1,200,574	156,877	13.2%
23	Córdoba	1,462,909	191,797	13.1%	1,555,596	102,251	6.6%	1,784,783	139,543	7.8%
25	Cundinamarca	2,228,682	73,517	3.3%	2,792,877	12,884	0.5%	2,919,060	22,358	0.8%
27	Chocó	388,476	285,964	73.6%	457,412	337,442	73.8%	534,826	418,970	79.5%
41	Huila	1,001,476	11,516	1.1%	1,009,548	5,027	0.5%	1,100,386	12,455	1.1%
44	La Guajira	655,943	91,674	14.0%	825,364	60,256	7.3%	880,560	76,050	8.5%
47	Magdalena	1,136,819	110,186	9.7%	1,263,788	106,143	8.4%	1,341,746	120,011	9.8%
50	Meta	713,772	17,911	2.5%	919,129	8,710	0.9%	1,039,722	8,867	1.6%
52	Nariño	1,498,234	270,433	18.1%	1,335,521	232,847	17.4%	1,630,592	369,192	23.6%
54	Norte de Santander	1,208,336	22,022	1.8%	1,346,806	5,410	0.4%	1,491,689	15,266	1.0%
63	Quindío	518,691	12,718	2.5%	509,640	6,036	1.2%	539,904	9,300	1.7%
66	Risaralda	859,666	43,503	5.1%	839,597	16,608	2.0%	943,401	24,613	2.7%
68	Santander	1,913,444	59,707	3.1%	2,008,841	22,554	1.1%	2,184,837	40,600	1.8%
70	Sucre	762,263	121,624	16.0%	864,036	102,655	11.9%	904,863	110,541	12.2%
73	Tolima	1,312,304	15,766	1.2%	1,228,763	5,110	0.4%	1,330,187	14,145	1.1%
76	Valle del Cauca	4,052,535	1,090,943	26.9%	3,789,874	646,762	17.1%	4,475,886	742,879	17.0%
81	Arauca	153,028	5,905	3.9%	239,503	9,994	4.2%	262,174	10,286	4.9%
85	Casanare	281,294	3,971	1.4%	379,892	6,089	1.6%	420,504	6,216	2.7%
86	Putumayo	237,197	11,620	4.9%	283,197	10,220	3.6%	348,182	10,582	5.2%
88	Archipiélago de San Andrés	59,573	10,452	17.5%	48,299	6,531	13.5%	61,280	10,766	18.3%
91	Amazonas	46,950	866	1.8%	66,056	473	0.7%	76,589	476	5.6%
94	Guainía	18,797	184	1.0%	44,431	451	1.0%	48,114	459	4.1%
95	Guaviare	56,758	2,883	5.1%	73,081	2,980	4.1%	82,767	3,136	6.4%
97	Vaupés	19,943	270	1.4%	37,690	270	0.7%	40,797	271	3.8%
99	Vichada	44,592	1,114	2.5%	76,642	570	0.7%	107,808	577	9.5%
Total		41,468,384	4,273,722	10.3%	44,164,417	2,950,072	6.7%	48,258,494	3,599,207	7.5%

Distribución proporcional mediante tabla cuadra a nivel departamental y municipal

Fase IV



**El futuro
es de todos**

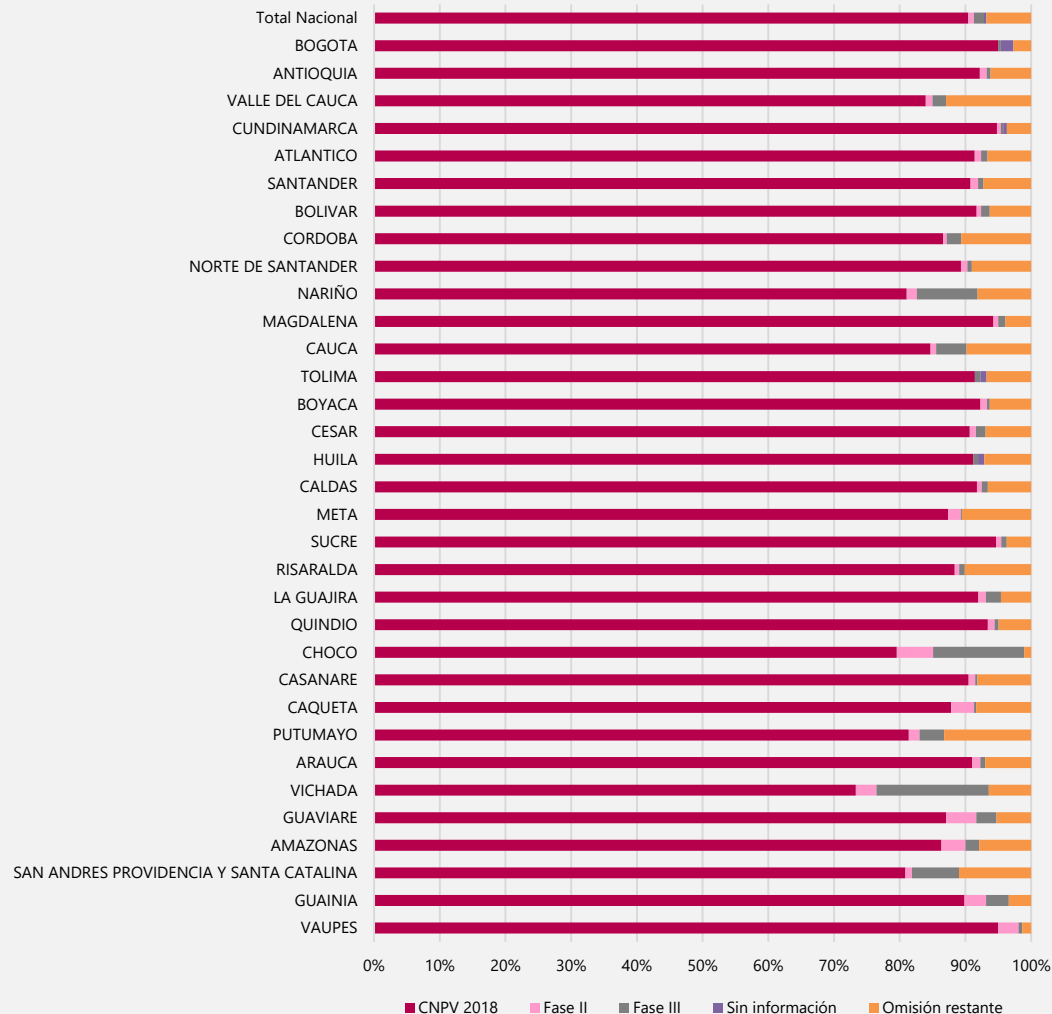
**Gobierno
de Colombia**



El objetivo de esta fase, es armonizar las estimaciones de fase II y III mediante la distribución proporcional de la población omitida que no fue asignada en las anteriores fases.

Supuestos:

1. Distribuir la población omitida sin asignar, mediante una tabla cuadrada a nivel departamental.
2. Tomar como referente las participaciones étnicas dadas por la ECV 2018 a nivel nacional – Condición Columna.
3. Tomar como referente el total de poblacional ajustada por cobertura – Condición Fila.
4. Condicionar el procedimiento para no estimar un total poblacional inferior al inferior a lo efectivamente censado.
5. Una vez se tenga convergencia, aplicar el mismo procedimiento a nivel municipal dentro del departamento.





El insumo para la tabla cuadrada departamental contiene la población efectivamente censada, estimación de fase II, estimación de fase III, población ajustada por cobertura y el remanente de omisión no asignado por las fases anteriores.

Departamento	Indígena	Gitano(a) o Rrom	Raizal del Archipiélago de San Andrés, Providencia y Santa Catalina	Palenquero(a) de San Basilio	Negro(a), mulato(a), afrodescendiente, afrocolombiano(a)	Ningún grupo étnico	Sin información de pertenencia étnica	Restante Omisión CNPV 2018	P. BASE CNPV 2018
	4,34%	0,01%	0,04%	0,02%	9,28%	86,31%			
Total Nacional	2.036.604	2.649	25.515	6.637	3.599.207	39.196.596	182.484	3.391.286	48.258.494
ANTIOQUIA	38.811	140	640	183	348.095	5.622.424	0	396.809	6.407.102
ATLANTICO	39.061	101	478	852	163.282	2.162.330	0	169.413	2.535.517
BOGOTÁ	19.131	603	1.060	218	97.541	7.094.869	142.564	199.144	7.412.566
BOLIVAR	5.305	31	573	3.988	342.289	1.584.174	0	133.750	2.070.110
BOYACA	7.409	18	62	16	9.176	1.124.181	0	76.514	1.217.376
CALDAS	57.730	37	107	30	21.983	852.659	0	65.709	998.255
CAQUETA	11.372	21	30	14	5.112	345.570	0	39.730	401.849
CAUCA	338.442	39	93	86	287.295	684.503	0	154.030	1.464.488
CESAR	55.057	20	128	75	156.877	903.421	0	84.996	1.200.574
CORDOBA	205.483	142	167	77	139.543	1.249.124	0	190.247	1.784.783
CUNDINAMARCA	19.621	98	148	60	22.358	2.769.734	16.820	107.041	2.919.060
CHOCO	81.683	36	130	124	418.970	21.861	0	12.022	534.826
HUILA	17.547	35	43	29	12.455	992.220	10.334	78.057	1.100.386
LA GUAJIRA	399.841	29	108	111	76.050	365.276	0	39.145	880.560
MAGDALENA	21.869	39	95	80	120.011	1.135.841	0	63.811	1.341.746
META	24.613	32	96	30	8.867	889.526	0	116.558	1.039.722
NARIÑO	234.231	141	114	101	369.192	879.096	0	147.717	1.630.592
NORTE DE SANTANDER	4.841	238	38	22	15.266	1.336.452	0	134.832	1.491.689
QUINDIO	3.120	6	22	2	9.300	500.638	0	26.816	539.904
RISARALDA	29.935	18	96	29	24.613	792.687	0	96.023	943.401
SANTANDER	1.412	347	156	49	40.600	1.984.470	0	157.803	2.184.837
SUCRE	104.890	134	135	46	110.541	655.677	0	33.440	904.863
TOLIMA	47.312	161	60	37	14.145	1.178.126	12.766	90.346	1.330.187
VALLE DEL CAUCA	32.638	136	474	290	742.879	3.105.998	0	593.471	4.475.886
ARAUCA	8.308	4	51	13	10.286	222.846	0	20.666	262.174
CASANARE	8.590	12	28	13	6.216	366.770	0	38.875	420.504
PUTUMAYO	63.236	17	12	30	10.582	221.721	0	52.584	348.182
SAN ANDRÉS PROVIDENCIA Y SANTA CATALINA	1.871	0	20.332	10	10.766	21.392	0	6.909	61.280
AMAZONAS	40.904	5	13	0	476	25.584	0	9.607	76.589
GUAINIA	35.068	5	5	4	459	9.446	0	3.127	48.114
GUAVIARE	10.228	3	6	5	3.136	63.024	0	6.365	82.767
VAUPES	33.108	0	10	8	271	5.582	0	1.818	40.797
VICHADA	61.709	1	5	5	577	29.374	0	16.137	107.808

Del resultado de la tabla cuadrada se destacan el departamento del atlántico que sobrepasa las 200 mil personas autorreconocidas como NMAA y el departamento del Valle del Cauca que no supera el millón de personas NMAA.

Cada fila puede ser modelada mediante un distribución multinomial con 6 categorías, de la misma forma las columnas con 33 categorías.

Con esta aproximación es posible generar una distribución empírica mediante un Bootstrap paramétrico con 1000 replicas por fila y columna, del la cual se obtienen intervalos del 95% de confianza, tomando el percentil 2,5 y 97,5 de la distribución.

Departamento	Indígena	Rrom	Raizal	Palenquer o	NMAA	Ningún	Sin información de pertenencia étnica	Total
Total Nacional	2.134.859	2.692	25.537	11.106	4.396.885	41.504.751	182.664	48.258.494
Antioquia	39.635	140	642	322	427.813	5.938.550	0	6.407.102
Atlántico	40.017	102	480	1.493	204.185	2.289.240	0	2.535.517
Bogotá, D.C.	19.131	603	1.060	370	118.671	7.130.167	142.564	7.412.566
Bolívar	5.344	31	574	6.492	418.965	1.638.704	0	2.070.110
Boyacá	7.402	18	62	29	11.607	1.198.258	0	1.217.376
Caldas	57.942	37	107	54	27.900	912.215	0	998.255
Caquetá	11.585	21	31	26	6.729	383.457	0	401.849
Cauca	349.099	40	95	143	355.008	760.103	0	1.464.488
Cesar	55.623	20	129	132	193.927	950.743	0	1.200.574
Córdoba	216.780	150	169	144	183.311	1.384.229	0	1.784.783
Cundinamarca	17.508	98	148	105	27.520	2.856.843	16.838	2.919.060
Chocó	90.510	36	130	167	422.011	21.972	0	534.826
Huila	17.322	36	43	52	15.898	1.056.636	10.399	1.100.386
La Guajira	406.991	29	109	193	93.446	379.792	0	880.560
Magdalena	22.316	39	95	137	145.629	1.173.530	0	1.341.746
Meta	25.110	33	97	57	11.850	1.002.575	0	1.039.722
Nariño	240.182	147	118	162	434.360	955.623	0	1.630.592
Norte de Santander	4.915	245	38	41	19.876	1.466.574	0	1.491.689
Quindio	3.121	6	22	3	11.573	525.179	0	539.904
Risaralda	30.517	20	96	53	32.422	880.293	0	943.401
Santander	1.407	351	157	87	51.727	2.131.108	0	2.184.837
Sucre	105.937	136	137	76	126.708	671.869	0	904.863
Tolima	47.518	161	60	66	18.015	1.251.504	12.863	1.330.187
Valle del Cauca	33.396	143	474	535	980.440	3.460.898	0	4.475.886
Arauca	8.309	4	51	22	13.175	240.613	0	262.174
Casanare	8.737	12	28	25	8.114	403.588	0	420.504
Putumayo	69.132	19	12	61	14.929	264.029	0	348.182
Archipiélago de San Andrés	1.933	0	20.332	18	14.530	24.467	0	61.280
Amazonas	45.173	5	14	0	679	30.718	0	76.589
Guainía	36.913	6	5	7	618	10.565	0	48.114
Guaviare	10.228	3	6	9	4.042	68.479	0	82.767
Vaupés	34.147	0	10	15	363	6.262	0	40.797
Vichada	70.979	1	6	10	844	35.968	0	107.808

Intervalos de Confianza por Fila





Intervalos de Confianza por Columna





Conclusiones FASE I y II

1. Los diseños muestrales de las encuestas y sus resultados demográficos se encuentran dentro del estándar de calidad estadístico, además muestran un comportamiento estructural semejante al CNPV 2018.
2. Las probabilidades condicionales permiten asumir que existe una fuerte relación entre los fonemas de los nombres, lugar de nacimiento o el lugar de residencia de las personas con el autorreconocimiento étnico.
3. Aunque las regiones de Amazonia, Bogotá y Orinoquia no han alcanzado el diagnóstico mínimo para predecir de forma correcta, se da por concluida la predicción en la región Pacífica. La región Andina y Atlántica alcanzan niveles aceptables en la calidad del ajuste y se podría aplicar su predicción, sin embargo se está explorando la partición de dichas regiones con el fin de tomar el mejor modelo viable.
4. Ajustar los modelos con todo el conjunto de datos aumenta la precisión, sin embargo las matrices de confusión muestran una disminución en el poder predictivo del modelo.



Conclusiones FASE III

1. La segmentación temática mejora de forma significativa el desempeño predictivo de los modelos empleados.
2. Con un coeficiente de determinación de **98%, 89% y 92%** para los grupos de regiones 1, 2 y 3, respectivamente se evidencia que el uso de indicadores censales explican más de 90% de la variabilidad en la participación NMAA de los conjuntos de aprendizaje.
3. Las métricas de RMSE son inferiores al 5%.
5. La participación NMAA a nivel nacional pasa de un 6,7% a un 7,5%, aumentando la población en **601.764**.
6. El grupo 1 compuesto por las regiones Atlántica, Pacífica y Suroccidente muestra el mejor desempeño, ya que cuenta con las mayores participaciones de la población NMAA.
7. Se destaca que los modelos no sobrepasan la estimación del total nacional de población NMAA, sin embargo la diferencia es cercana al millón de personas respecto a lo estimado por ECV.
8. Aplicando la proporción NMAA observada dentro de las manzanas a las personas omitidas, se alcanza una participación nacional de **6,8%**.



Conclusiones FASE IV

1. La implementación de la tabla cuadrada consigue armonizar las estimaciones realizadas en fase II y III de forma parsimoniosa.
2. La estimación de población NMAA en total nacional cae dentro del intervalo de confianza de la ECV 2018.
3. Aunque la población NMAA en Valle del Cauca no es cercana a lo estimado en la ECV 2018, la propuesta metodológica da cuenta de una participación del 22% dentro del departamento, así como del 22% del total nacional, siendo el departamento con mayo población NMAA en el territorio.
4. Los comprimentos de los intervalos de confianza son pequeños, lo que indica un alto grado de precisión en las estimaciones.
5. Se recomienda la publicación de estos resultados bajo el enfoque de estadísticas experimentales.
6. Queda abierta la posibilidad de encontrar estrategias que permitan determinar las personas que están en la categoría de ningún grupo étnico (No realización de la pregunta, marcación arbitraria por parte del encuestador o asignación errónea por parte del informante idóneo)



Bibliografía

- ◉ <https://spark.apache.org/docs/latest/ml-classification-regression.html>
- ◉ file:///systemasasnv/Direccion_Geoestadistica/AFRO/PRUEBA_290420/Metodologia_ECV_%202018.pdf
- ◉ file:///systemasasnv/Direccion_Geoestadistica/AFRO/PRUEBA_290420/Metodologia_General_Gran_Encuesta_Integrada_de_Hogares.pdf
- ◉ Voicu, I. (2018). *Using First Name Information to Improve Race and. Statistics and Public Policy*, 5:1, 1-13, DOI: 10.1080/2330443X.2018.1427012.
- ◉ Stevens, F. R. (2015). *Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data*. PLOS ONE, 1-22.

Estimación de población étnica a niveles subnacionales

Aprovechamiento de información auxiliar



**El futuro
es de todos**

**Gobierno
de Colombia**



@DANEColombia



@DANE_Colombia



/DANEColombia



/DANEColombia

www.dane.gov.co