¿Quién va a vender su voto?
Predicción de clientelismo en Colombia

David Santiago Caraballo Candela *
Sergio David Pinilla Padilla †
Juan Diego Valencia Romero ‡

12 de diciembre de 2022

Resumen

¿Es posible predecir la susceptibilidad de los ciudadanos a practicar el clientelismo en las elecciones? El intercambio de votos por beneficios particularistas es esencial para comprender la relación entre la sociedad civil y el Estado. La práctica de acciones clientelistas establece una estructura de incentivos para políticos y ciudadanos que es perjudicial para el desarrollo de la capacidad estatal y el funcionamiento de la democracia. Este trabajo persigue un objetivo concreto: determinar quién venderá su voto en las próximas elecciones. Introducimos una novedosa estrategia de predicción clientelista basada en datos originales de Colombia. Nuestra métrica de evaluación corresponde a la minimización de una función de pérdida ajustada *W-MIR* que pondera asimétricamente los cuadrados de la tasa de falsos negativos (2/3) y falsos positivos (1/3). Encontramos que los árboles de decisión funcionan mejor de acuerdo con nuestra intención; en particular, la especificación *XGBoost*. Sin embargo, los modelos mejor equilibrados son los que incorporan estadística Bayesiana.

Palabras clave: Clientelismo, elecciones, métrica, predicción, parámetros, falsos negativos, falsos positivos.
Códigos JEL: C55, D72, P37.

GitHub: Repositorio Final Entrega

*Universidad de los Andes, Facultad de Economía. ds.caraballo@uniandes.edu.co.
†Universidad de los Andes, Facultad de Economía. sd.pinilla@uniandes.edu.co.
‡Universidad de los Andes, Facultad de Economía. jd.valenciar@uniandes.edu.co.

1 Introducción

¿Se puede predecir la susceptibilidad de los ciudadanos a practicar el clientelismo en elecciones? El clientelismo electoral es la entrega condicionada de beneficios particulares a un individuo o grupo de personas a cambio de su voto en periodos electorales. Este fenómeno es clave para entender la naturaleza del electorado y las instituciones democráticas así como la relación entre la sociedad civil y el Estado (Fergusson et al., 2020). Existe un consenso general sobre la prevalencia de estas conductas que apuntan hacia el detrimento de la democracia ya que socava los vínculos más pragmáticos entre los ciudadanos y los políticos. Por lo tanto, el comportamiento intrínseco del electorado pertenece al conjunto de decisiones económicas basadas en el análisis costo-beneficio. Como los beneficios de una acción clientelar superan sus costos, hay incentivos para ejecutarla. La primera forma de aproximar estos costos es en términos del bienestar general. Una acción clientelista parece inofensiva en el corto plazo, pero con el tiempo, el intercambio del voto por dinero en efectivo, u otras formas de coimas, en lugar de la provisión bienes y servicios públicos resulta perniciosa para la sociedad (Bates, 1981; Kitshelt, 2000).

Desde la perspectiva de la economía política, el término del clientelismo ha demostrado ser adaptable a diversos entornos políticos, sociales y culturales. Su misma expresión sirve para denominar sistemas, instituciones o individuos nocivos para el funcionamiento normativo de la democracia. Hicken (2011) argumenta que si bien no existe una definición ampliamente aceptada, varias de ellas destacan elementos claves de las relaciones clientelares cómo diádicas, de contingencia, de jerarquía e iteración[1]. Los políticos del gobierno reparten "barril de tocino", *pork-barrel* como se conoce despectivamente en la literatura, cuando gastan un alto nivel de recursos en sus distritos electorales para animar a los votantes a (re)elegirlos (Grossman & Helpman, 2005). La idea que subyace a esta práctica es que el dinero asignado al distrito del representante "cambiará positivamente" la vida de los electores locales, asegurando así su apoyo -presente o futuro- y sus votos.

Las prácticas clientelistas son un fenómeno extendido en democracias en desarrollo, particularmente en América Latina, y ampliamente documentado bajo distintas perspectivas metodológicas y analíticas. La literatura relacionada coincide en que la "venta" del voto, entre otras cosas: entorpece la rendición de cuentas y suprime la responsabilidad de quienes ocupan un cargo de elección (Stokes, 2005); coopta la habilidad ciudadana para elegir libremente a los gobernantes que los identifica, mina la calidad de la política pública que emana de aquellos que ganaron con prácticas clientelistas (Greene, 2016); y agudiza las desigualdades sociales con mayor intensidad y frecuencia en localidades pobres (Sandholt & Justesen, 2013; Vicente, 2014) toda vez que genera un gasto desmesurado cuyo origen de los recursos es, mayormente, producto del delito (e.g. narcotráfico).

[1]En el orden en que aparecen. (1) Interacciones y transacciones cara a cara entre el patrón y el cliente. (2) La entrega de un bien o servicio por parte de cliente en respuesta directa a la entrega de un beneficio recíproco de la contraparte, o a la promesa creíble de dicho beneficio. (3) Una relación en la que un individuo de mayor estatus socio-económico (patrón) utiliza su propia influencia y recursos para proporcionar protección y/o beneficios a una persona de menor estatus (cliente). (4) Cada parte anticipa futuras interacciones mientras toman decisiones sobre su comportamiento actual.

En contraste con la poca atención que ha recibido este problema en otras latitudes, en Colombia continúa siendo un asunto pendiente en la agenda académica. A esta también se añade la prevalencia de aproximaciones cuantitativas, en donde se prioriza el uso de datos socio-demográficos y de encuestas (Fergusson et al., 2017). Examinarlo empíricamente es un reto de diseño. El clientelismo supone un obstáculo para la investigación experimental en las ciencias sociales por la dificultad de obtener respuestas sinceras en las encuestas, especialmente en temas sensibles para el electorado como la discriminación, la corrupción y la ilegalidad. Los encuestados pueden proveer respuestas falsas cuando se les pregunta, y la naturaleza del comportamiento implica que hay pocas alternativas de recuperarlo. Así, al identificar con mayor precisión a potenciales individuos que practicarían el clientelismo, es posible trazar estrategias de control y prevención que mejoren la transparencia de los procesos electorales y frenen la erosión de la democracia (Gonzalez-Ocantos et al., 2012).

La amplia gama de trabajos empíricos se centran en la provisión de correlaciones y efectos causales para estudiar los distintos aspectos del clientelismo[2] dentro de colectivos de individuos. Esto se debe, presuntamente, la academia ha considerado al clientelismo en Colombia como una característica agregada del electorado, entendible únicamente en niveles locales, departamentales o regionales. Este trabajo introduce una estrategia de predicción que persigue un objetivo particular: determinar quién va a vender su voto en una contienda electoral. Nuestra aproximación empírica se basa en el entrenamiento, calibración e implementación de un conjunto de modelos estadísticos y de aprendizaje de maquinas[3] sobre una base de datos representativa a nivel nacional de características socio-demográficas, laborales y electorales registradas en 2016. Además, realizamos un ejercicio de estática comparativa entre los modelos para medir su desempeño de acuerdo con una métrica de evaluación -cuya creación es propia-, de tal forma que determine cuál es la mejor estrategia de predicción.

Los resultados sugieren que un modelo *XGBoost* es el que mejor anticipa si un colombiano incurrirá en conductas clientelistas durante las elecciones. El desempeño del modelo supera nuestros *benchmarks*, incluyendo el que se obtendría si se asigna la clase de manera aleatoria a los individuos de manera que replicara la proporción de clientelistas sobre el total de personas en la base ($\approx 13\%$). Al mismo tiempo, el *XGBoost* es uno de los que menor frecuencia presenta en la *False Positive Rate (FPR)*. En efecto, se le considera que este es el que mejor se adapta al objetivo del trabajo. Cabe resaltar que el *QDA* corresponde a un *second-best* razonable que, si bien no presenta el mejor desempeño en la clasificación de clientelistas, si incurre en una baja *False Negative Rate (FNR)*.

[2]Entre ellas, la discriminación de los afroamericanos y otros grupos minoritarios o marginados, actitudes hacia la comida, comportamientos sexuales de riesgo y otras acciones delicadas o ilegales. Los experimentos con listas se han utilizado también para estudiar el comportamiento electoral (Holbrook & Krosnick, 2010; Corstange, 2010), como hacemos en este trabajo. Recientemente se han utilizado en Colombia para estudiar el apoyo a determinados grupos, especialmente militares y rebeldes (Matanock & García-Sánchez, 2011; Steele & Shapiro, 2012), así como el clientelismo en García & Pantoja (2015).

[3]Para un total de ocho (8) modelos clasificatorios distintos: i) *K-Nearest Neighbors (KNN)*; ii) *Linear Discriminant Analisys (LDA)*; iii) *Quadratic Discriminant Analisys (QDA)*; iv) *Elastic-Net*; v) *Random Forest*; vi) *XGBoost*; vii) *LightGBM*, y; vi) *Artificial Neural Network*.

Más allá de estas dos metodologías, las demás se aglutinan en dos grupos: i) modelos no paramétricos que clasifican consistentemente a los que sí venden su voto -baja *FNR*-, pero que incurren en una alta *FPR* y ii) modelos lineales con baja *FPR*, pero que son propensos a identificar correctamente a quienes sí practican el clientelismo -alta *FNR*-.

Nuestro trabajo se relaciona con dos vertientes de la literatura. Primero, siguiendo a González (2020), contribuye a la literatura de economía política al abordar la susceptibilidad de los ciudadanos a practicar el clientelismo por medio de un modelo de clasificación. Segundo, contribuye a la literatura de aprendizaje de máquinas en economía aplicada al combinar técnicas de predicción avanzadas y compararlas con métricas de creación propia para la predicción de un fenómeno político. Incluyendo la introducción, este documento cuenta con cuatro secciones. La sección 2 describe los datos que se utilizaron para el ejercicio. En la sección 3 se describe y analiza el desempeño de los modelos que se entrenaron junto con las métricas de evaluación propuestas. Finalmente, la sección 4 presenta las conclusiones del ejercicio y otras anotaciones.

2 Datos

Los datos sobre los cuales se entrenaron y evaluaron los modelos provienen de la ola 2016 de la Encuesta Longitudinal Colombiana de la Universidad de los Andes (Bernal et al., 2014), o ELCA. La ELCA es la primera encuesta de hogares que es de tipo panel a gran escala en Colombia. Esta cuenta con aproximadamente 10.000 hogares representativos de la zona urbana, junto con cinco macro-regiones rurales. La línea base se realizó en 2010. En el 2013, el primer seguimiento incluyó un Módulo de Política (MP) aplicado a un miembro del hogar (el jefe del hogar o su pareja/cónyuge, asignado aleatoriamente cuando ambos se encontraban disponibles). El nuevo MP abarcó preguntas sobre participación e interés en la política, fuentes de información acerca de las noticias del país, posiciones ideológicas y, fundamentalmente, una pregunta para estudiar el intercambio de estos por beneficios, puestos de trabajo o regalos materiales. Esta ultima es nuestra variable dependiente, utilizada como *proxy* de clientelismo, pregunta explícitamente si:

> *"(...) para decidir por quién votar ¿usted ha tenido en cuenta beneficios, regalos o trabajos que un candidato le ofreció a usted o un familiar a cambio de su voto?"*

La base contiene cerca de 1.000 predictores a nivel de hogar o persona, los cuales capturan información acerca de los comportamientos y actitudes políticas de las personas, junto con la caracterización del hogar en términos ocupantes, el material de la vivienda y el estrato socio-económico, entre otras. Además, se tiene la información de los choques y riesgos que ha sufrido el hogar, a nivel de persona se cuenta con preguntas relacionadas con la salud, su capital social, su situación laboral, su educación y gastos, entre otras. Se seleccionaron 137 regresores para emplear en el entrenamiento de los modelos siguiendo dos criterios: i) se incluyeron aquellas con mayor potencial predictivo para el clientelismo que menciona la literatura, i.e. medidas de ingreso y educación de las personas, y; ii) se incluyeron otras que presentaban una correlación significativa con la variable de clientelismo pero que no se habían tenido en cuenta anteriormente.

La Tabla 1 (Anexos) profundiza en las estadísticas descriptivas de los individuos y hogares. El 40,78% de personas en la muestra son hombres, mientras que el 59,22% restante es mujer, con una edad media general de 48,85 años. El 31% de los encuestados en la base de datos se benefician del subsidio de Familias en Acción, mientras que un 80,15% participaron en la ultima elección local. Existen 7.900 personas mayores de edad (>18 años) para las cuales se tiene registro de la variable dependiente. Esta variable posee desbalance ya que únicamente el 13,16% de las personas en toda la base respondieron [1 =Sí] a la pregunta sobre actitudes clientelistas. En las zonas urbanas, el 13,09% de las personas revelaron su comportamiento clientelista, en las zonas rurales este valor se incrementó ligeramente hasta el 13,24%. La Figura 1 (Anexos) muestra que la región del Atlántico presenta la tasa más alta, con una prevalencia del 23,28% en las áreas urbanas y del 30,62% en el área rural. En contraste, la región Cundiboyacense es la que menor acciones clientelares reporta, con el 7,77%. El grado educativo con mayor frecuencia alcanzado corresponde al rango de sin estudios a básica secundaria (85,15%), el 14,85% restante respondieron haber recibido educación de nivel técnico, universitario y/o posgrado. Estos valores son consistente con la distribución de los estratos socioeconómicos en tanto los estratos 1, 2 y 3 representan cerca del 97% de la muestra. Adicionalmente, el 53,44% posee vivienda propia, 20,78% viven en arriendo y el 20,38% reside en una vivienda con permiso.

3 Una estrategia de predicción clientelista

La presente sección describe la metodología para determinar la ocurrencia de una acción clientelista a partir de las observaciones en la tercera ronda (2016) de la ELCA. Para poder evaluar el desempeño de los nueve (9) modelos de aprendizaje de maquinas implementados se van a utilizar tres criterios de información derivados de las matrices de confusión para ejercicios de clasificación sobre una variable dicótoma. El primero de estos corresponde al *FNR*, que para este caso muestra la proporción de clientelistas que fueron mal clasificados como no clientelistas, situación que es poco deseable según el objetivo de este trabajo. El segundo corresponde al *FPR*, que estaría mostrando cual es la proporción de no clientelistas que fueron erradamente clasificados como clientelistas, lo que tampoco es muy deseable, pero no va directamente en contra de los objetivos de este trabajo.

En este orden de ideas, teniendo en cuenta que para este trabajo debería ser más costoso incurrir en falsos negativos que en falsos positivos, como ultimo criterio de información se terminó utilizando una métrica de evaluación de elaboración propia, el *W-MIR*, que corresponde a una combinación lineal ponderada entre el *FNR* y el *FPR*, con pesos de dos a uno (2:1), respectivamente, y reescalado para que el máximo valor posible (cuando todas las observaciones están mal clasificadas) siga siendo de uno. Previo a comenzar a estimar cualquier modelo, vale la pena resaltar que, así como se menciono previamente, las categorías de la variable de interés están muy desbalanceadas dentro de la base de datos de la ELCA. Dado que tan solo aproximadamente el 13% de los encuestados afirmaron tener comportamientos clientelistas. Entonces, para poder realizar la evaluación

fuera de muestra de los modelos, se realizó un *train-test split* con una relación 80:20, y estratificado para preservar la proporción de clientelistas dentro de ambas submuestras.

Para solucionar el problema que el desbalance puede inducir sobre la capacidad de los modelos de identificar personas dispuestas a vender su voto, a causa de que estas están sub-representadas en la muestra, se utilizó una técnica de *over-sampling* sintético conocida como *SMOTE*[4] que balancea perfectamente a ambas categorías. Esto permitió que incrementara la información disponible para entrenar a los modelos con los atributos presentes en los individuos clientelistas, sin comprometer y/o disminuir la cantidad de información disponible en la misma muestra acerca de los atributos de los no clientelistas. Antes de calibrar los ocho (8) modelos más complejos, se decidió utilizar dos estrategias de clasificación que desempeñaron el rol de *baselines* para evaluar que tanto la complejidad de esos modelos adicionales servía para mejorar la identificación efectiva de clientelistas.

En principio, a partir de 15.000 iteraciones se realizó una asignación aleatoria de predicciones de clientelismo utilizando la misma proporción de verdaderos clientelistas en la muestra, lo cual llevó a un *W-MIR* por fuera de muestra de 0,631. Así, cualquier modelo que presente un *W-MIR*$< 0,631$ se podrá considerar como mejor en la predicción del clientelismo que una asignación aleatoria. En segundo lugar, como modelo *baseline* mediante máxima verosimilitud se estimó uno de los modelos clasificatorios más sencillos, un Logit sin regularización. Entonces, con un *W-MIR* por fuera de muestra de 0,412, esto implica que: i) si es posible presentar predicciones que sean mejores a las asignaciones aleatorias a un bajo costo computacional, y; ii) además, para que sea razonable utilizar cualquier modelo más complejo que un Logit en esta labor de clasificación, el modelo debería presentar un *W-MIR* por fuera de muestra significativamente inferior a 0,412.

Tabla 2: Métricas de evaluación: modelos de clasificación.

Metodología	*W-MIR*		*W-MIR Adj.*		*FNR*		*FPR*	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
Aleatorio	0,625	0,631	0,508	0,508	0,868	0,868	0,132	0,132
Logit	0,316	0,412	0,103	0,184	0,357	0,495	0,233	0,246
KNN	0,000	0,321	0,000	0,165	0,000	0,144	0,000	0,674
LDA	0,338	0,365	0,119	0,141	0,387	0,428	0,240	0,241
QDA	0,237	0,346	0,086	0,135	0,115	0,260	0,481	0,520
Elastic-Net	0,349	0,364	0,129	0,142	0,412	0,433	0,221	0,225
Random Forest	0,264	0,324	0,162	0,177	0,049	0,135	0,694	0,704
XGBoost	0,204	0,315	0,122	0,151	0,003	0,154	0,605	0,638
LightGBM	0,277	0,312	0,231	0,236	0,000	0,048	0,832	0,839
Artifical Neural Network	0,273	0,365	0,075	0,134	0,263	0,389	0,294	0,316

En la Tabla 2 se encuentran los resultados de las métricas de evaluación para las dos

[4] *Synthetic Minority Oversampling Technique:* Funciona seleccionando ejemplos cercanos en el espacio de características, trazando una línea entre los ejemplos en el espacio de características y dibujando una nueva muestra en un punto a lo largo de esa línea.

líneas bases y los ocho (8) modelos estudiados. Las columnas *Train* muestran los valores de las funciones de pérdida dentro de muestra, las columnas *Test* fuera de muestra.

Estos modelos provienen de vertientes de la literatura de aprendizaje de maquinas, la mayoría poseen métodos no paramétricos de clasificación[5], algunos también incluyen estrategias de selección de variables y reducción de coeficientes[6], otros incorporan estrategias de clasificación provenientes de la estadística Bayesiana[7], e inclusive aquellos diseñados a partir de árboles de decisión[8]. De esta manera, se está garantizando que el problema de clasificación de individuos clientelistas se esta abordando desde una amplia diversidad de metodologías, cada una de las cuales utiliza distintas herramientas estadísticas para intentar reducir el sesgo y/o la varianza de las predicciones. Para la calibración de los hiper-parámetros de cada uno de estos modelos, se utilizó inicialmente una búsqueda exhaustiva con *5-Fold Cross-Validation* a través del método `GridSearchCV`. Para evitar problemas de *overfitting*, se realizó una búsqueda manual de los hiper-parámetros con el mejor desempeño fuera de muestra, dentro de la vecindad de los mejores parámetros encontrados a partir de la `GridSearchCV`. A partir del uso de la *W-MIR* fuera de muestra evidenciamos que los modelos, ya calibrados, que mejor cumplían con el objetivo de este documento son el *LightGBM*, el *XGBoost*, y el *KNN*.

Sin embargo, a partir del uso de un *W-MIR* ajustado[9] que penaliza por el desbalance en las predicciones (ver Tabla 2), se concluye que de aquelos modelos que mejor predicen, el *XGBoost* es el que predice de forma mas armoniosa, lo cual implica que este es el modelo que mejor identifica a personas clientelistas mientras que al mismo tiempo se controla no clasificar a no clientelistas como clientelistas. Cabe resaltar que durante la calibración de este modelo hicimos especial énfasis en evitar potenciales problemas de *overfitting*, por lo que este modelo cuenta con 35 arboles, con una máxima profundidad de 3 niveles, y se le otorga un elevado peso a la clase minoritaria (1:16,5). En segundo lugar, también evidenciamos que dentro de los modelos con predicciones fuera de muestra mas armoniosas (*Artificial Neural Network*, *QDA* y *LDA*), el que posee mayor *W-MIR* es la especificación *QDA*, por lo tanto este correspondería a un modelo *second-best* que no tiene un mal desempeño predictivo (puede llegar a cumplir con el objetivo de este trabajo), pero que adicionalmente si es muy bueno a la hora de procurar que no esta clasificando mal ni a los clientelistas ni a los no clientelistas.

El único parámetro calibrado durante la estimación del *QDA* fue el regularizador, $\alpha^* = 0,425$, el cual permitió controlar la varianza de las predicciones de este modelo. En línea con lo anterior, se presentó una tendencia interesante entre los dos grupos de modelos a los que estos pertenecen. Por un lado, los modelos derivados de los arboles de decisión presentan un menor sesgo porque se desempeñan mucho mejor que los demás a la hora de clasificar bien la categoría que se desea predecir bien. No obstante, tienden a

[5] *KNN*, *Random Forest*, *XGBoost*, *LightGBM*, *Artificial Neural Network*.
[6] *Elastic-Net*, *Random Forest*, *XGBoost*, *LightGBM*
[7] *LDA*, *QDA*, *Elastic-Net*
[8] *Random Forest*, *XGBoost*, *LightGBM*
[9] *W-MIR Ad.*= $\frac{1}{3} \times FPR^2 + \frac{2}{3} \times FNR^2$

estar sujetos a mayor varianza, ya que se sobre-ajustan a las características propias de cada observación, conllevando a que, tal como se aprecia en la Tabla 2, el *W-MIR* se eleve mucho cuando se pasa de las predicciones dentro del *Train* al *Test*.

Por otro lado, los modelos derivados de estadística Bayesiana puede que presenten un mayor sesgo, dado que no son tan buenos prediciendo bien la pertenencia a una categoría en especifico (ej. si una persona tendrá comportamientos clientelistas). Pero, gracias a la presencia de *priors* implícitos y de términos de reducción de los coeficientes, su varianza es menor y presentan un *W-MIR* consistente entre *Train* y *Test*, debido a que este no se ven tan afectado por cambios ligeros en las características de los individuos.

4 Conclusiones y otras anotaciones

La venta del voto evita que las demandas ciudadanas sean incorporadas en la agenda política, en especial de los más necesitados, quienes son el principal objetivo de los intermediarios. Esta compra de votos favorece a quienes acceden al poder por esta vía y perjudica a los estratos más bajos, pero son estos últimos la principal fuente de sufragios. Partidos y candidatos continúan tomando ventaja de los pobres y de su falta de autonomía. Este trabajo se encontró que, a partir de los datos de 2016 de la ELCA, el mejor modelo para predecir si un individuo colombiano va a vender su voto, dándole mayor ponderación a predecir bien la incidencia de clientelismo que a predecir la no incidencia de clientelismo, corresponde a un *XGBoost* con alrededor de 35 arboles y con poca profundidad, para minimizar el *overfitting*. Adicionalmente, se esperaría que este modelo no sea tan desbalanceado en sus predicciones como otros modelos también provenientes de arboles de decisión. En particular, esto puede ocurrir debido a que a diferencia de sus pares el mismo *XGBoost* tiende a penalizar el proceso de optimización y la complejidad de sus ramas, lo cual permitiría contrarrestar aun mas buena parte del sesgo presente en los modelos de arboles de decisión, mejorando así el balance de sus predicciones.

Sin embargo, también vale la pena resaltar que como este modelo presenta una mayor varianza, tal vez su uso no seria tan recomendable sobre poblaciones muy diferentes a la que fue entrenado o en periodos de tiempo distintos al 2016 (en este caso). Si, tal vez, se esta dispuesto a sacrificar algo de presición por una menor varianza y una mayor estabilidad en las predicciones se podría hacer uso de un *QDA* con un parámetro de regularización de alrededor de $\alpha^* = 0,425$.

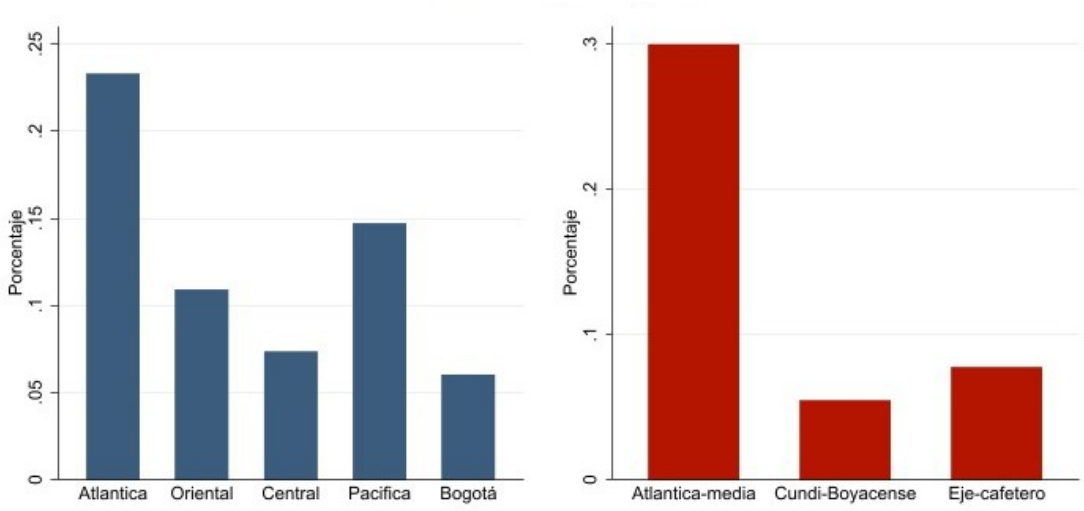
Se resaltan varias limitaciones del ejercicio. En primer lugar, el tamaño de la muestra (7.900) puede no ser suficiente para entrenar de manera adecuada modelos complejos que requieren de una gran disponibilidad de datos, como las redes neuronales o el *XGBoost*, por lo que no es posible determinar si el desempeño de estos modelos podría sobrepasar al de los demás en escenarios con mayor disponibilidad de datos. Lo anterior es importante, pues el desempeño fuera de muestra de los modelos no superó de manera tan holgada a la clasificación ingenua de todos los individuos como clientelistas ($W - MIR = 0,33$). Es posible que esto mejore en con una base mas grande, sin embargo, la disponibilidad

de bases con preguntas relacionadas con el clientelismo es todavía escasea en Colombia.

Anexos

Figuras

Figura 1: Incidencia del clientelismo por regiones.



Tablas

Tabla 1: Estadísticas descriptivas.

VARIABLE	Obs.	Media	Desviación	Mediana
Sexo [1=Hombre]	7900	0,40	0,49	0
Edad	7900	48,85	12,54	49
Clientelista [1=Sí]	7900	0,13	0,34	0
Sindicato [1=Pertence]	7572	0,01	0,11	0
Familias en Acción [1=Recibe]	7900	0,31	0,46	0
Voto Alcaldía [1=Votó]	7900	0,80	0,40	1
Zona [1=Urbano]	7900	0,55	0,50	1
Sin estudios - Básica secundaria	7900	0,85	0,36	1
Técnico y tecnológico	7900	0,13	0,33	0
Universitaria y superior	7900	0,02	0,13	0
Posgrado	7900	0,01	0,08	0
Estrato 1	7900	0,42	0,49	0
Estrato 2	7900	0,42	0,49	0
Estrato 3	7900	0,13	0,33	0
Estrato 4	7900	0,02	0,14	0
Estrato 5 y 6	7900	0,01	0,04	0
Vivienda propia	7900	0,53	0,50	1
Vivienda en pago	7900	0,05	0,21	0
Vivienda en arriendo	7900	0,21	0,41	0
Vivienda con permiso	7900	0,20	0,40	0
Vivienda sin título	7900	0,01	0,08	0

Referencias

Bates, R. (1981). Markets and states in tropical Africa: the political basis of agricultural policies. *Berkeley: University of California Press*.

Bernal, R., Cadena, X., Camacho, A., Cárdenas, J., Fergusson, L., & Ibañez, A. (2014). Encuesta Longitudinal de la Universidad de los Andes (ELCA)–2013. *Documentos CEDE*.

Corstange, D. (2010). Vote buying under competition and monopsony: evidence from a list experiment in Lebanon. *Paper prepared for the 2010 Annual Conference of the American Political Science Association, Washington, D.C.*.

Fergusson, L., Molina, C., & Riaño, J. (2017). I sell my vote, and so what? A new database and evidence from Colombia. *Documentos CEDE*, ISSN 1657-7191.

Fergusson, L., Molina, C., & Robinson, J. (2020). The Weak State Trap. *National Bureau of Economic Research*.

García, M., & Pantoja, S. (2015). Incidencia del clientelismo según riesgo electoral y de violencia. *Misión de Observación Electoral*.

Gonzalez-Ocantos, E., de Jonge, C., Meléndez, O. J., C., & Nickerson, D. (2012). Vote buying and social desirability bias: experimental evidence from Nicaragua. *American Political Science Review*, 56 (1), pp. 202-217.

González, T. (2020). Compra de voto en Colombia: ¿cómo viste el fantasma y cuáles son sus implicaciones? *Reflexión Política*, 22(46), 44–57. (Recuperado de: https://doi.org/10.29375/01240781.3992)

Greene, K. (2016). Why Vote Buying Fails: Campaign Effects and the Elusive Swing Voter. *University of Texas*.

Grossman, G., & Helpman, E. (2005). Party Discipline and Pork-barrel Politics. *National Bureau of Economic Research*, NBER Working Paper Series.

Hicken, A. (2011). Clientelism. *Department of Political Science, University of Michigan*, 14 (1), pp. 289-310.

Holbrook, A., & Krosnick, J. (2010). Social desirability bias in voter turnout reports: tests using item count technique. *Public Opinion Quarterly*, 74 (1), pp. 37-67.

Kitshelt, H. (2000). Linkages between citizens and politicians in democratic politics. *Documentos CEDE*, 33 (6-7), pp. 845-79.

Matanock, A., & García-Sánchez, M. (2011). Fighting for hearts and minds: examining popular support for the military in Colombia. *Presented at the Governance, Development, and Political Violence Conference at the University of California at San Diego*.

Sandholt, P., & Justesen, M. (2013). Poverty and Vote Buying: Survey-based evidence from Africa. *Electoral Studies*, 33, pp. 220-232. (Recuperado de: http://doi.org/10.W16jelectstud.2013.07.020)

Steele, A., & Shapiro, J. (2012). State-builing, ounterinsurgency, and development in Colombia. *Unpublished paper*.

Stokes, S. (2005). Perverse Accountability: A Formal Model of Machine Politics with Evidence from Argentina. *American Political Science Review*, 99 (3), pp. 315-325. (Recuperado de: http://doi.org/10.1017/S0003055405051683)

Vicente, P. (2014). Is Vote Buying Effective? Evidence from a Field Experiment in West Africa. *The Economic Journal*, 124(574), pp. 356-387. (Recuperado de: http://doi.org/10.1111/ecoj.12086)