Determinantes del Voto Republicano en las Elecciones de 2020

Álvaro Pérez

31 de mayo 2024

Resumen

Este estudio analiza los determinantes del voto por el Partido Republicano en las elecciones presidenciales de 2020 en Estados Unidos, enfocándose en la influencia de la inmigración mexicana. Utilizando modelos de regresión múltiple, logística y random forest, se examinan datos electorales y demográficos a nivel de condado para evaluar cómo la presencia de población mexicana y otros factores socioeconómicos influyen en el comportamiento electoral.

Palabras Clave: elecciones, determinantes del voto, voto republicano, comportamiento electoral, retórica anti-inmigrante

Código JEL: D72

1. Introducción

Este trabajo de investigación es parte de un primer avance y aproximación a un trabajo más robusto con miras a convertirse, enventualmente, en una tesis de Ciencia Política del Instituto Tecnológico Autónomo de México. Es el resultado del trabajo de un semestre cursando el Seminario de Investigación Política con énfasis en relaciones ejecutivo-legislativo. Decidí enfocarme en un tema electoral y de la política estadounidense con perspectiva de migración. El tema de investigación es el estudio de la elección presidencial de 2020 en Estados Unidos y el efecto de los mexicanos en el voto republicano.

Para evaluar el impacto de la inmigración mexicana en el comportamiento electoral, se emplearon modelos estadísticos econométricos. Estos modelos permiten examinar la relación entre la proporción de población mexicana en un condado y el porcentaje de votos por el Partido Republicano, así como la influencia de otros factores socioeconómicos y demográficos. Estimar la relación entre la cantidad de migrantes que habitan en un condado y el número de votos recibidos por políticos que impulsan medidas antiinmigración permite ahondar en la investigación para determinar qué tan significativo es el impacto de la inmigración sobre el voto. Es decir si un aumento en la cantidad de inmigrantes realmente puede alterar las preferencias electorales o más bien son otras preocupaciones, como la ansiedad económica o la política exterior, las que influyen en la decisión de los votantes.

La siguiente sección hace una revisión de la literatura existente y plantea las hipótesis. A continuación se describen los datos empleados. La cuarta sección presenta los tres modelos usados, tanto su formalización matemática como su adecuación al presente caso de estudio. La quinta sección muestra los resultados para cada uno de los modelos junto con sus respectivas interpretaciones. La última sección concluye y explora futuras investigaciones.

2. Marco Teórico

La campaña de Donald Trump en 2016 y 2020 incluyó una fuerte retórica anti-inmigrante, que resonó con ciertos segmentos del electorado. Esta retórica afectó de manera notable a los condados con alta población de inmigrantes mexicanos. Estudios previos han encontrado que la percepción de una amenaza económica y cultural debido a la inmigración puede movilizar a los votantes nativos

a favor de candidatos que promueven políticas restrictivas de inmigración.

El análisis de Goetz et al. (2018) indicó que los condados con mayores proporciones de población hispana, y particularmente mexicana, tendieron a votar menos por Trump en comparación con otros condados. Esto se debe a que las políticas propuestas por Trump fueron percibidas como directamente perjudiciales para estas comunidades. Sin embargo, en algunos casos, la presencia de una población inmigrante significativa puede movilizar a los votantes nativos a favor de candidatos que perciben como defensores de políticas de control migratorio.

El estudio de Flaxman et al. (2021) utilizó datos del censo y resultados electorales para realizar inferencias ecológicas y entender las tendencias de votación a nivel local. Encontraron que la presencia de inmigrantes y la diversidad étnica están correlacionadas con patrones específicos de votación. En áreas con alta inmigración mexicana, la amenaza percibida por los residentes nativos puede llevar a un mayor apoyo a candidatos con políticas anti-inmigrantes, como Trump. Descubrieron que las variables demográficas como la raza, la educación y los ingresos fueron predictivas del voto por Trump. Sus resultados sugieren que el nivel educativo y la identidad racial fueron factores determinantes en las decisiones de voto.

El factor migración y la presencia de mexicanos en los condados de Estados Unidos tiene un impacto significativo en la dinámica del voto republicano. La percepción de amenaza y la retórica anti-inmigrante han jugado un papel crucial en movilizar a ciertos segmentos del electorado a favor de candidatos como Donald Trump.

Las hipótesis del presente trabajo de investigación son las siguientes:

1. **H1:** La retórica anti-inmigrante moviliza el voto republicano.

   En los condados con una alta proporción de población mexicana, la retórica anti-inmigrante de los candidatos republicanos está positivamente asociada con un mayor porcentaje de votos por el Partido Republicano.

2. **H2:** El impacto de la población mexicana sobre el voto republicano es no lineal.

   La relación entre la proporción de población mexicana y el porcentaje de votos por el Partido Republicano es positiva hasta cierto punto, después del cual comienza a disminuir.

3. **H3:** Factores socioeconómicos y demográficos moderan el impacto de la migración en el voto republicano.

La relación entre la proporción de población mexicana y el voto republicano es más fuerte en condados con menores niveles de educación. La tasa de pobreza en un condado moderará la relación entre la proporción de población mexicana y el voto republicano, de modo que en condados más pobres, la influencia de la población mexicana será más pronunciada.

A través de un análisis detallado de los datos electorales y demográficos a nivel de condado, se espera proporcionar una comprensión más profunda de cómo la presencia de mexicanos influye en las dinámicas del voto republicano.

3. Datos

Los datos obtenidos para esta investigación provienen principalmente de dos fuentes. Las variables electorales, o sea todos los datos que son resultados de elecciones presidenciales provienen del MIT Election Lab. La base de datos contiene el número de votos por cada candidatura presidencial desde 2020 a nivel condado. Para mis modelos, usé el porcentaje de voto efectivo de cada partido, eliminando los votos nulos y de terceros candidatos. En el caso de los resultados de la elección presidencial de 1980, extraje la información del Dave Leip's Atlas of U.S. Presidential Elections. La variable repsh80 la uso como control en los modelos.

Los datos de las variables socioeconómicas en su mayoría fueron obtenidos del Census Bureau. Aquellas variables que estaban en términos absolutos las convertí a términos relativos, es decir, como porcentaje del total. La variable Ingreso per Cápita (ppi) la obtuve del Bureau of Economic Analysis y la variable Religión, del The Association of Religious Data Archives.

3.1. Análisis exploratorio de datos

El Cuadro 1 presenta un resumen descriptivo de los datos a nivel de condado, incluyendo la media, desviación estándar, valor mínimo y valor máximo para cada variable. Las variables analizadas incluyen porcentajes de votos por el Partido Republicano y Democrático, características demográficas como el porcentaje de mexicanos, personas con título de grado y tasa de empleo, y otras variables socioeconómicas como el coeficiente de Gini, tasa de pobreza, ingreso per cápita y el porcentaje de hogares que reciben asistencia gubernamental. Además, se incluyen diferencias en el porcentaje de votos republicanos entre elecciones. Esta descripción proporciona una visión general de las distribuciones y variabilidad de las variables clave utilizadas en el estudio.

Cuadro 1: Descripción y EDA de los datos a nivel de condado

Variable	Descripción	Media	Desv. est.	Min.	Max.
repsh20	% votos por el Partido Republicano en 2020	0.667	0.150	0.055	0.945
depsh20	% votos por el Partido Demócrata en 2020	0.332	0.150	0.054	0.944
mex	% mexicanos por condado	0.048	0.082	0.002	0.908
edu	% personas con título de grado	0.146	0.053	0.004	0.381
emp	Tasa de empleo	0.586	0.074	0.217	0.789
pov	% personas por debajo de la línea de pobreza	0.141	0.058	0.020	0.589
gin	Coeficiente de Gini	0.443	0.036	0.291	0.696
asis	% hogares que reciben asistencia gubernamental	0.022	0.013	0.000	0.268
rel	% católicos por condado	0.109	0.106	0.0005	0.957
ppi	Ingreso percápita	$47,843	11394.47	$ 22,265	$ 300,665
rur	% personas que viven en áreas rurales	0.6322	0.316	0.000	1.00
rpsh80	% voto por el Partido Republicano en 1980	0.577	0.136	0.151	0.90
dif1612	diferencia del % voto por el Partido Republicano en 2016 menos 2012	0.068	0.161	-0.665	0.724
dif2016	diferencia del % voto por el Partido Republicano en 2020 menos 2016	0.169	2.12	-0.668	0.701

3.2. Correlaciones

La Figura 1 presenta una matriz de correlación que ayuda a identificar visualmente las variables más relacionadas. Los círculos indican la fuerza y dirección de la correlación entre las variables, donde el tamaño del círculo corresponde a la magnitud de la correlación y el color indica la dirección (positivo en azul y negativo en rojo).

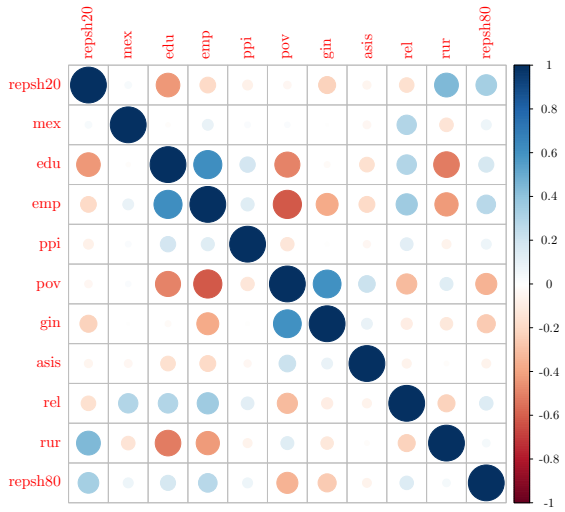

Figura 1: Matriz de correlaciones

4. Metodología

4.1. Modelo de Regresión Múltiple

En esta investigación, se utilizó un modelo de regresión lineal múltiple para analizar los factores que influyen en el porcentaje de voto por el Partido Republicano en las elecciones de 2020. La variable dependiente (Y) es el porcentaje de voto por el Partido Republicano (`repsh2020`), y las variables independientes (X_i) incluyen diversas características demográficas, económicas y sociales.

Se estimaron tres modelos de regresión lineal múltiple con la variable dependiente repsh2020 (porcentaje de voto por el Partido Republicano). Los modelos se especificaron de la siguiente manera:

$$\begin{aligned}
\text{Modelo 1 (M1)} : \text{repsh2020} = {} & \beta_0 + \beta_1 \text{mex} + \beta_2 \text{mex2} + \beta_3 \text{edu} + \beta_4 \text{emp} \\
& + \beta_5 \text{ppi} + \beta_6 \text{pov} + \beta_7 \text{gin} + \beta_8 \text{asis} \\
& + \beta_9 \text{rel} + \beta_{10} \text{repsh16} + \beta_{11} \text{repsh80} + \beta_{12} \text{rur} + \epsilon \qquad (1) \\
\text{Modelo 2 (M2)} : \text{repsh2020} = {} & \beta_0 + \beta_1 \text{mex} + \beta_2 \text{mex2} + \beta_3 \text{edu} + \beta_4 \text{emp} \\
& + \beta_5 \text{ppi} + \beta_6 \text{pov} + \beta_7 \text{gin} + \beta_8 \text{asis} \\
& + \beta_9 \text{rel} + \beta_{10} \text{repsh16} + \beta_{11} \text{repsh80} + \beta_{12} \exp(\text{rur}) + \epsilon \qquad (2) \\
\text{Modelo 3 (M3)} : \text{repsh2020} = {} & \beta_0 + \beta_1 \text{mex} + \beta_2 \text{edu} + \beta_3 \text{emp} + \beta_4 \text{ppi} \\
& + \beta_5 \text{pov} + \beta_6 \text{gin} + \beta_7 \text{asis} + \beta_8 \text{rel} + \beta_9 \text{rur} + \epsilon \qquad (3)
\end{aligned}$$

Donde:

- Y es el porcentaje de voto por el Partido Republicano en 2020 (`repsh2020`).
- β_0 es el término constante o intercepto.
- β_i son los coeficientes de regresión para cada una de las variables independientes.
- ϵ es el término de error.

Los datos utilizados en este análisis provienen de diversas fuentes, incluyendo censos y encuestas nacionales. Se realizó una limpieza y transformación de los datos para garantizar su adecuación al

modelo de regresión. Las variables `mex` y `ppi` fueron elevadas al cuadrado para capturar posibles relaciones no lineales.

El modelo se estimó utilizando el método de mínimos cuadrados ordinarios (OLS, por sus siglas en inglés). Este método permite encontrar los coeficientes β_i que minimizan la suma de los cuadrados de los residuos, proporcionando la mejor estimación lineal de la relación entre las variables independientes y la variable dependiente.

4.2. Modelo de Regresión Logística

La regresión logística es una técnica adecuada para modelar variables dependientes binarias, en este caso, la proporción de votos para el Partido Republicano en 2020. La variable `repsh20` es una variable dummy que se define como:

$$\text{repsh20} = \begin{cases} 0, & \text{si repsh20} < 0{,}5 \\ 1, & \text{si repsh20} > 0{,}5 \end{cases}$$

El modelo se especifica de la siguiente manera:

$$\begin{aligned} \text{Logit}(P(\text{repsh20} = 1)) = \beta_0 &+ \beta_1 \text{mex} + \beta_2 \text{edu} + \beta_3 \text{emp} \\ &+ \beta_4 \text{ppi} + \beta_5 \text{pov} + \beta_6 \text{gin} \\ &+ \beta_7 \text{asis} + \beta_8 \text{rel} + \beta_9 \text{repsh16} \\ &+ \beta_{10} \text{repsh80} + \beta_{11} \text{rur} \end{aligned} \tag{4}$$

Donde:

- $\text{Logit}(P(\text{repsh20} = 1))$ es el logit de la probabilidad de que un condado vote por el Partido Republicano en 2020.
- β_0 es el intercepto del modelo.
- β_i son los coeficientes de las variables independientes.

4.3. Modelo Random Forest

El modelo Random Forest es un método de aprendizaje supervisado que combina múltiples árboles de decisión para mejorar la precisión y controlar el sobreajuste. Este modelo es particularmente útil para manejar grandes conjuntos de datos con múltiples variables independientes. Se basa en el concepto de ensamblado (ensemble learning), que combina múltiples modelos para mejorar el rendimiento y la robustez en comparación con un solo modelo. La formalización matemática es como sigue:

Dado un conjunto de datos de entrenamiento $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$:

1. Para $b = 1$ hasta B:
   - Se genera una muestra bootstrap D_b de tamaño n a partir de D.
   - Se construye un árbol de decisión T_b utilizando D_b y seleccionando aleatoriamente m características en cada división.
2. La predicción para una nueva observación x se realiza de la siguiente manera:
   - Clasificación:
   $$\hat{y} = \text{mode}\{T_b(x)\} \quad \text{para} \quad b = 1, 2, \ldots, B.$$
   - Regresión:
   $$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x).$$

Al tener muchas variables independientes y desear un modelo que ayude a hacer predicciones precisas y entender la importancia de cada variable, es conveniente usar random forest. La lógica del modelo es que si se está tratando de predecir el porcentaje de voto por el Partido Republicano en un condado específico, en lugar de confiar en un solo árbol de decisión que podría decir que el porcentaje de población mexicana es lo más importante, al usar random forest se crean 100 árboles de decisión. Un árbol podría informar que el nivel educativo y la tasa de empleo son los factores más importantes. Otro árbol podría enfocarse más en el ingreso personal per cápita. Al final, el random forest toma todas estas «opiniones» y las combina para dar una predicción promedio, que es más confiable que cualquier predicción individual. Así, el modelo da una medida de la importancia de cada variable.

5. Resultados

5.1. Regresión Multivariada

En esta sección se presentan los resultados del modelo de regresión lineal múltiple estimado para analizar los factores que influyen en el porcentaje de voto por el Partido Republicano en las elecciones de 2020.Los coeficientes estimados, errores estándar, valores t y niveles de significancia para cada una de las variables independientes se presentan en el Cuadro 2.

La proporción de población mexicana se asocia positivamente con el porcentaje de voto por el Partido Republicano en todos los modelos. Este efecto es significativo en todos los modelos. Por otra parte, el coeficiente negativo para el cuadrado de la proporción de población mexicana indica una relación no lineal, sugiriendo que el efecto positivo de la población mexicana disminuye a mayores niveles de dicha población. Este efecto es significativo en todos los modelos.La Figura 2 respalda esta hallazgo. A medida que aumenta la proporción de mexicanos por condado aumenta el voto republicano, hasta que se rebasa cierto punto (alrededor del 60 % de mexicanos), empieza a disminuir el voto republicano.

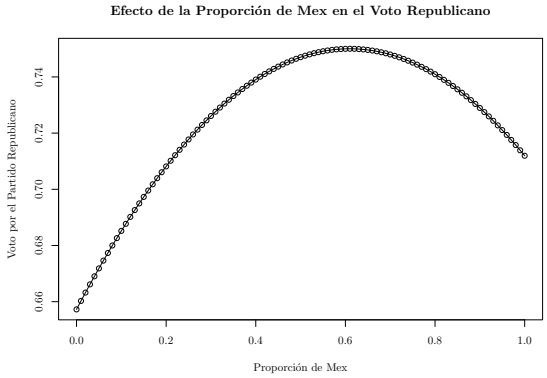

Figura 2: Efecto de la Proporción de Mex en la Predicción Ajustada

En cuanto a la educación, un mayor nivel educativo promedio de la población se asocia consistentemente con una disminución en el porcentaje de voto por el Partido Republicano en todos los modelos. Este efecto es altamente significativo. La tasa de empleo tiene un efecto negativo y significativo en los Modelos 1 y 2, pero no es significativo en el Modelo 3. El ingreso per cápita tiene un coeficiente muy pequeño pero significativo en todos los modelos, indicando que un mayor

Cuadro 2: Modelos de Regresión

	Variable dependiente: voto por el Partido Republicano		
	Modelos		
	(M1)	(M2)	(M3)
mex	0.318***	0.304***	0.184***
	(0.027)	(0.027)	(0.014)
mex2	−0.510***	−0.496***	
	(0.053)	(0.053)	
edu	−1.226***	−1.270***	−1.191***
	(0.027)	(0.027)	(0.031)
emp	−0.126***	−0.140***	−0.004
	(0.020)	(0.020)	(0.023)
ppi	−0.00000***	−0.00000***	−0.00000*
	(0.00000)	(0.00000)	(0.00000)
pov	−0.608***	−0.622***	−0.747***
	(0.028)	(0.028)	(0.031)
gin	−0.062*	−0.079**	−0.166***
	(0.037)	(0.037)	(0.043)
asis	−0.404***	−0.438***	−0.632***
	(0.075)	(0.075)	(0.086)
rel	−0.120***	−0.123***	−0.156***
	(0.010)	(0.010)	(0.012)
repsh16	0.242***	0.241***	
	(0.007)	(0.007)	
repsh80	0.378***	0.380***	
	(0.009)	(0.009)	
rur	0.084***		0.116***
	(0.004)		(0.004)
exp(rur)		0.039***	
		(0.002)	
Constant	0.624***	0.625***	0.980***
	(0.022)	(0.023)	(0.025)
Observaciones	12,481	12,481	12,482
R^2	0.502	0.498	0.343
R^2 Ajustada	0.502	0.498	0.342
Error Est. Residual	0.106 (df = 12468)	0.106 (df = 12468)	0.122 (df = 12472)
Estadístico F	1,047.332*** (df = 12; 12468)	1,032.023*** (df = 12; 12468)	721.913*** (df = 9; 12472)
Nota:			*p<0.1; **p<0.05; ***p<0.01

ingreso se asocia con una ligera disminución en el porcentaje de voto por el Partido Republicano.

La tasa de pobreza tiene un efecto negativo y altamente significativo en todos los modelos, sugiriendo que una mayor pobreza se asocia con una disminución en el porcentaje de voto por el Partido Republicano. Lo que resulta muy contraintuitivo, pues se esperaría que en entornos más pobres (ligados a menor empleo y menos alfabetismo, por mencionar algunas) se observaría un alto porcentaje de voto por el Partido Republicano. Igualmente interesante resulta el coeficiente asociado al índice de Gini, éste es negativo y significativo en los Modelos 2 y 3, lo que indica que una mayor desigualdad económica se asocia con una disminución en el porcentaje de voto por el Partido Republicano. Mientras que una mayor proporción de la población que recibe asistencia social se asocia con una disminución en el porcentaje de voto por el Partido Republicano. Este efecto es altamente significativo en todos los modelos.

La proporción de la población católica se asocia negativamente con el porcentaje de voto por el Partido Republicano en todos los modelos, con alta significancia. Habría que realizar el análisis con otras denominaciones cristianas e inclusive otras religiones. En el caso de los Modelos 1 y 3 una mayor proporción de la población rural se asocia positivamente con el porcentaje de voto por el Partido Republicano, con alta significancia. En el Modelo 2 esta variable es exponencial, pero igualmente significativa. Como era de esperarse, el porcentaje de voto por el Partido Republicano en 2016 y en 1980 se asocia positivamente con el porcentaje de voto en 2020.

Los Modelos 1 y 2 explican aproximadamente el 50 % de la variabilidad en el porcentaje de voto por el Partido Republicano, mientras que el Modelo 3 explica aproximadamente el 34 %. Esto indica que los Modelos 1 y 2 tienen un mejor ajuste que el Modelo 3. Además, los Modelos 1 y 2 tienen un error estándar residual de 0.106, mientras que el Modelo 3 tiene un error estándar residual de 0.122, esto sustenta que los Modelos 1 y 2 proporcionan un ajuste más preciso. Todos los modelos son estadísticamente significativos en su conjunto ($p < 0.001$), con el Modelo 1 presentando el mayor valor de F, seguido por el Modelo 2 y finalmente el Modelo 3.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
12468	139.99				
12468	141.03	-0	-1.03		

Cuadro 3: Prueba F para la comparación de modelos

El Modelo 1 y 2 son muy parecidos, pero para decidir cuál de los dos es mejor empleo una prueba F (Cuadro 3). El resultado de la prueba F para comparar los Modelos 1 y 2 no muestra

una mejora significativa en el ajuste del Modelo 2 sobre el Modelo 1. Esto se refleja en el valor de Sum of Sq = -1.0348 y el hecho de que no hay un valor de F y Pr(>F) significativo que indique que el Modelo 2 es mejor que el Modelo 1. Esto sugiere que añadir la variable exp(rur) no mejora significativamente el ajuste del modelo. Adicionalmente, usé los criterios de información de Akaike (AIC) y Bayesiano (BIC). Estos criterios penalizan los modelos más complejos, por tanto, valores más bajos indican un mejor modelo. En ambos criterios el Modelo 1 obtuvo valores más bajos. Por tanto, tomando en cuenta el R^2 ajustado, ligeramente mayor el del Modelo 1; la Prueba F, que no mostró una mejora significativa al incluir la variable adicional en el Modelo 2; y los valores de AIC y BIC más bajos para el Modelo 1, podemos concluir que el Modelo 1 es superior al Modelo 2.

df	AIC
M1	-20596.97
M2	-20505.05

Cuadro 4: Criterio de Información de Akaike (AIC)

df	BIC
M1	-20492.92
M2	-20401.01

Cuadro 5: Criterio de Información Bayesiano (BIC)

5.2. Regresión Logística

Básicamente en este modelo se encontró que un aumento en casi todas las variables independientes está asociado con una disminución en la probabilidad de votar por el Partido Republicano, salvo por la variable de proporción de población rural y claro, repsh16 y repsh80. Hace sentido que a mayor educación, mayor ingreso, mayor empleo disminuya la probabilidad de votar republicano, lo que sintoniza con que entre más población rural haya más voto republicano se observa. El cambio de signo entre mex y mex2 se explico en el apartado anterior. Lo interesante es que a mayor pobreza y mayor desigualdad disminuya la probabilidad de votar republicano.

Para evaluar el modelo empleo la curva ROC. La curva ROC es una herramienta gráfica que se utiliza para evaluar la capacidad de un modelo de clasificación binaria. En el eje y (ordenadas) se encuentra la sensibilidad (verdaderos positivos) y en el eje x (abscisas) se encuentra la especificidad (1 - falsos positivos). La curva ROC presentada en la Figura 3 muestra que el modelo de regresión logística tiene una buena capacidad de discriminación. La curva se eleva rápidamente hacia la esquina superior izquierda, lo que indica una alta sensibilidad y especificidad.

La línea diagonal gris representa una clasificación aleatoria (AUC = 0.5). Cuanto más lejos esté la curva ROC de esta línea, mejor es el modelo en la clasificación de las observaciones. El área

	Dependent variable:
	Voto por el Partido Republicano
mex	2.093^{**}
	(0.921)
mex2	-4.574^{**}
	(1.793)
edu	-24.281^{***}
	(0.900)
emp	-1.670^{**}
	(0.723)
ppi	-0.00001^{***}
	(0.00000)
pov	-13.942^{***}
	(0.889)
gin	-5.644^{***}
	(1.262)
asis	-10.870^{***}
	(2.354)
rel	-1.731^{***}
	(0.340)
repsh16	3.232^{***}
	(0.204)
repsh80	8.189^{***}
	(0.333)
rur	0.854^{***}
	(0.130)
Constant	5.307^{***}
	(0.770)
Observations	12,481
Log Likelihood	−3,321.606
Akaike Inf. Crit.	6,669.211
Note:	*p<0.1; **p<0.05; ***p<0.01

Cuadro 6: Modelo de Regresión Logística

bajo la curva (AUC) es una métrica que resume el rendimiento de la curva ROC en un solo valor. Un AUC de 0.5 indica un modelo sin capacidad de discriminación, equivalente a una clasificación aleatoria; mientras, que un AUC de 1 indica un modelo perfecto que clasifica todas las observaciones correctamente. En este caso, la AUC es de 0.8896, lo que sugiere que el modelo tiene una excelente capacidad para distinguir entre los condados que votaron por el Partido Republicano y los que no. Adicionalmente realicé una matriz de confusión. Las siguientes métricas fueron derivadas de la

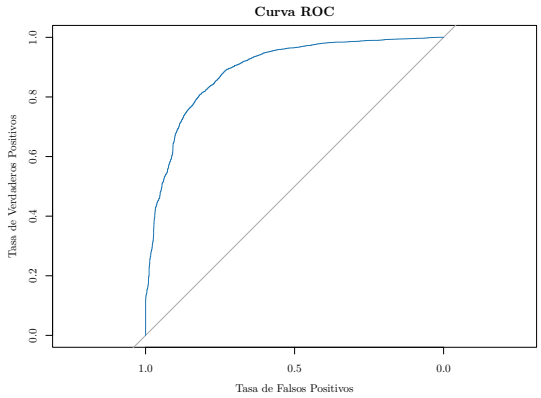

Figura 3: Curva ROC

matriz de confusión (Cuadro 7):

Referencia	0	1
Predicción		
0	869	316
1	985	10311

Cuadro 7: Matriz de Confusión

La exactitud es la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) sobre el total de predicciones. En este caso, el modelo tiene una exactitud del 89.58 %, lo que significa que el 89.58 % de las predicciones del modelo fueron correctas. La estadística kappa mide la concordancia entre las predicciones del modelo y los valores reales, ajustada por la concordancia que ocurre por azar. Un valor de 0.5158 indica una moderada concordancia.

La sensibilidad o tasa de verdaderos positivos, mide la proporción de positivos reales que son correctamente identificados por el modelo. Aquí, el 46.87 % de los condados que votaron por el Partido Republicano fueron correctamente clasificados como tales por el modelo. La especificidad

Métrica	Valor
Exactitud	0.8958
95 % IC	(0.8903, 0.9011)
No Information Rate	0.8515
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.5158
Prueba de Mcnemar P-Value	< 2.2e-16
Sensibilidad	0.46872
Especificidad	0.97026
Valor Pred Pos	0.73333
Valor Pred Neg	0.91280
Prevalencia	0.14855
Detection Rate	0.06963
Detection Prevalence	0.09494
Balanced Accuracy	0.71949

Cuadro 8: Estadísticas de la matriz de confusión

mide la proporción de negativos reales que son correctamente identificados por el modelo. En este caso, el 97.03 % de los condados que no votaron por el Partido Republicano fueron correctamente clasificados como tales por el modelo.

El valor predictivo positivo mide la proporción de predicciones positivas correctas. Aquí, el 73.33 % de los condados predichos como votantes por el Partido Republicano realmente votaron por dicho partido. Por otro lado, el valor predictivo negativo mide la proporción de predicciones negativas correctas. El 91.28 % de los condados predichos como no votantes por el Partido Republicano realmente no votaron por dicho partido.

La exactitud equilibrada es el promedio de la sensibilidad y la especificidad. Proporciona una medida más equilibrada del rendimiento del modelo cuando las clases están desbalanceadas. Aquí, la exactitud equilibrada es del 71.95 %. Finalmente, la prueba de McNemar evalúa si hay una diferencia significativa entre las proporciones de falsos positivos y falsos negativos. Un p-valor muy bajo indica que hay una diferencia significativa, sugiriendo que el modelo tiene un sesgo en su clasificación.

En conclusión este no es un muy buen modelo predictivo dado que sobreestima la cantidad de condados que votarían por el Partido Repúblicano. Habría que hacer ajustes a las variables de la regresión logística, pero después poder lograr un nivel de predicción más certera.

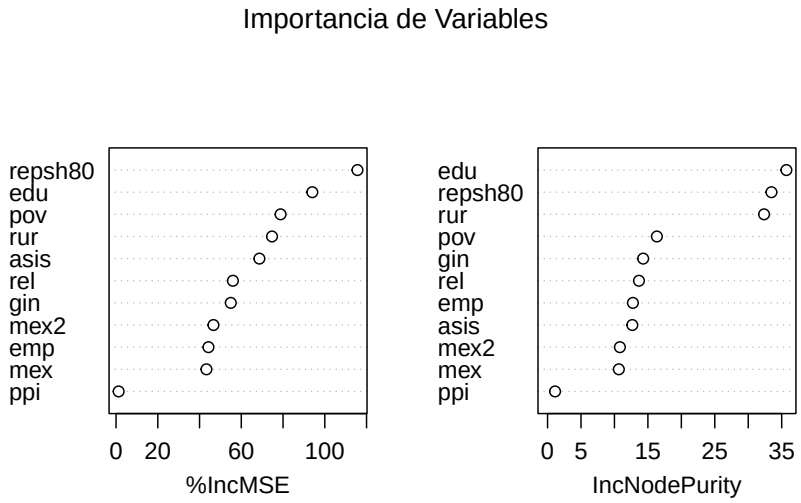

Figura 4: Importancia de Variables

5.3. Random Forest

Los datos se dividieron en un conjunto de entrenamiento (70 %) y un conjunto de prueba (30 %). Se entrenó el modelo con 500 árboles de decisión. El rendimiento del modelo se evaluó utilizando las siguientes métricas:

- Mean of Squared Residuals: 0.001556678
- Porcentaje de Varianza Explicada: 93.03 %

Estas métricas indican que el modelo Random Forest tiene una alta capacidad de predicción y explica el 93.03 % de la varianza en el porcentaje de voto por el Partido Republicano. El análisis de la importancia de las variables mostró que las más influyentes en la predicción del porcentaje de voto por el Partido Republicano fueron educación, ruralidad y pobreza (Figura 4).

El gráfico de la izquierda (Incremento en el Error Cuadrático Medio) muestra la importancia de las variables basada en el aumento porcentual del error cuadrático medio (MSE) cuando se permuta aleatoriamente una variable. Una mayor disminución en el MSE indica que la variable es más importante para el modelo. En otras palabras, si eliminar una variable causa un gran aumento en el error, esa variable es crucial para las predicciones del modelo.

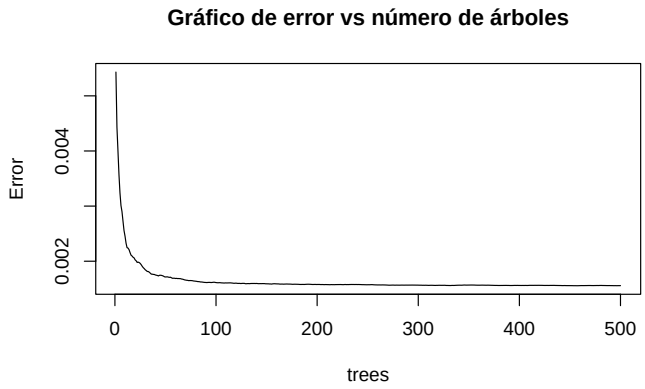

Figura 5: Error vs No. Árboles

- edu: El nivel educativo promedio también es un predictor más significativo excluyendo el voto en 1980.
- pov, rur, asis: Estas variables también contribuyen significativamente a la precisión del modelo.
- ppi: Aunque está en la parte inferior del gráfico, sigue siendo una variable importante, pero con menos impacto relativo en comparación con las otras.

El gráfico de la derecha (Incremento en la Pureza de los Nodos) mide la importancia de las variables basada en la reducción total de la impureza (en términos de varianza) en los nodos donde la variable se utiliza para hacer una división. Un mayor valor indica una mayor contribución a la división eficaz de los datos en los árboles de decisión.

- edu y repsh80: Estas dos variables son las más importantes según esta métrica, ya que generan las mayores reducciones en la impureza de los nodos. Esto refuerza la idea de que el nivel educativo y el historial de voto republicano son cruciales para la predicción.
- rur, pov: También son variables importantes que ayudan a dividir eficazmente los datos.
- ppi: Nuevamente, aunque está en la parte inferior, sigue contribuyendo a la pureza de los nodos, aunque en menor medida.

El gráfico de error vs el número de árboles (Figura 5) indica que cómo varía el error del modelo Random Forest a medida que se incrementa el número de árboles en el bosque. Al principio, cuando

el número de árboles es bajo, el error disminuye rápidamente. Esto se debe a que cada árbol adicional contribuye significativamente a mejorar la precisión del modelo. A medida que se añaden más árboles (alrededor de 100 árboles), la disminución del error se ralentiza y el error comienza a estabilizarse. En este punto, agregar más árboles tiene un impacto cada vez menor en la mejora del rendimiento del modelo. Finalmente, el error se estabiliza casi por completo después de unos 200 árboles.

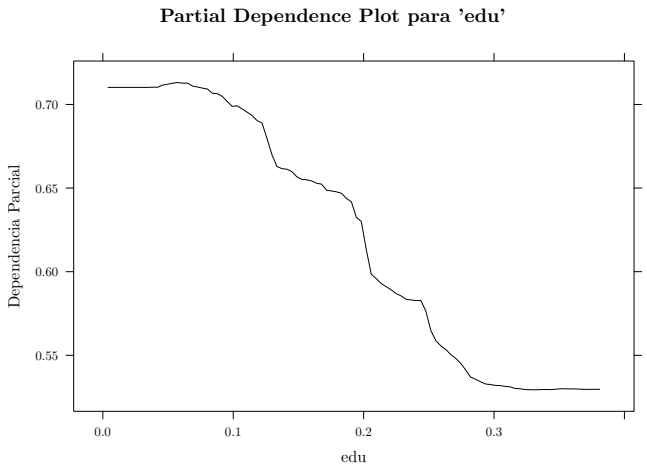

Figura 6: Partial Dependence Plot para 'edu'

Por último, el gráfico de dependencia pacial, PDP (Figura 6) muestra la relación marginal entre la variable de interés, que en este caso elegí edu (el nivel educativo promedio) y la variable de respuesta, denotada por yhat, (el porcentaje predicho de voto por el Partido Republicano) mientras se mantienen constantes todas las demás variables en el modelo.

A medida que aumenta el nivel educativo promedio (edu), el porcentaje predicho de voto por el Partido Republicano (yhat) disminuye. Esto sugiere que, en general, los condados con un nivel educativo promedio más alto tienden a tener un menor porcentaje de voto por el Partido Republicano. Hay una disminución notable en yhat cuando edu aumenta de alrededor de 0.05 a 0.15, este rango muestra que incluso pequeños aumentos en el nivel educativo tienen un impacto significativo en la reducción del voto republicano. A partir de un nivel educativo promedio de aproximadamente 0.25 en adelante, yhat se estabiliza cerca de 0.55, esto quiere decir que en condados con un nivel educativo promedio alto, el porcentaje de voto por el Partido Republicano se reduce y se estabiliza.

6. Conclusiones

Los resultados de esta investigación indican que múltiples factores demográficos, económicos y sociales tienen un impacto significativo en el porcentaje de voto por el Partido Republicano en las elecciones de 2020. Los modelos de regresión múltiple y logística sugieren que la proporción de población mexicana está positivamente asociada con el voto republicano hasta cierto punto, después del cual este efecto disminuye. Además, se encontró que variables como el nivel educativo y la tasa de pobreza moderan la influencia de la inmigración en el comportamiento electoral. El modelo random forest confirma la importancia de estos factores y ofrece una precisión robusta en las predicciones. Estos hallazgos tienen implicaciones importantes para el diseño de campañas electorales y estrategias de comunicación política, subrayando la necesidad de considerar el contexto local y las características demográficas al analizar los patrones de voto.

Con más datos se pueden perfeccionar los modelos para alcanzar coeficientes de determinación bastante elevados y con significancia estadística. Particularmente la regresión logística debe de ajustarse para que sus estimaciones sean más efectivas. Adicionalmente se pueden realizar otros modelos para elegir el mejor. Para continuar el trabajo de investigación se puede ahondar en cómo la relación entre la inmigración y el voto republicano ha evolucionado en elecciones anteriores y proyectar posibles tendencias futuras. Esto podría incluir un análisis de las elecciones de medio término y otras elecciones locales para obtener una visión más completa de las dinámicas electorales a lo largo del tiempo. Igual podría usar el método de diferencias-en-diferencias para aislar el efecto de los migrantes en un condado comparando el cambio en el voto por el Partido Republicano entre condados con similares niveles de migración antes de un periodo migratorio y después de éste.

La investigación tiene muchas áreas de oportunidad de mejora y desarrollo, que pueden seguir trabajándose y explorándose.

Referencias

[1] Bureau of Economic Analysis. *Regional Economic Accounts*. Accessed: 2024-05-29. 2023. URL: https://www.bea.gov.

[2] MIT Election Data y Science Lab. *County Presidential Election Returns 2000-2020*. Ver. V12. 2018. DOI: 10.7910/DVN/VOQCHQ. URL: https://doi.org/10.7910/DVN/VOQCHQ.

[3] Seth Flaxman et al. "Understanding the 2016 US Presidential Election using ecological inference and distribution regression with census microdata". En: *arXiv preprint arXiv:1611.03787* (2016).

[4] Manuel Diaz Garcia. "The Social Identities of Biden and Trump Sympathizers–A Multilevel-Analysis of the Candidate Preference in the US Presidential Election 2020". En: ().

[5] Stephan J Goetz et al. "Explaining the 2016 vote for president trump across us counties". En: *Applied Economic Perspectives and Policy* 41.4 (2019), págs. 703-722.

[6] Dave Leip. *Atlas of U.S. Presidential Elections: 1980*. 1980. URL: https://uselectionatlas.org/RESULTS/datagraph.php?year=1980.

[7] The Association of Religious Data Archives. *ARDA: American Religion Data Archive*. Accessed: 2024-05-29. 2023. URL: https://www.thearda.com.

[8] United States Census Bureau. *Data Access and Dissemination Systems (DADS)*. 2023. URL: https://www.census.gov.