



Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Estadística Aplicada

**DETERMINACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE DE MEJOR
AJUSTE PARA LA PREDICCIÓN DEL ABSTENCIONISMO ELECTORAL EN GUATEMALA**

Ing. Carlos Ottoniel Guevara Carranza

Asesorado por la Mtra. Inga Mayra Virginia Carvajal Castillo

Guatemala, noviembre de 2021

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA



FACULTAD DE INGENIERÍA

**DETERMINACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE DE MEJOR
AJUSTE PARA LA PREDICCIÓN DEL ABSTENCIONISMO ELECTORAL EN GUATEMALA**

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

ING. CARLOS OTTONIEL GUEVARA CARRANZA

ASESORADO POR LA MTRA. INGA. MAYRA VIRGINIA CARVAJAL CASTILLO

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN ESTADÍSTICA APLICADA

GUATEMALA, NOVIEMBRE DE 2021

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA



NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Cruz Lorente
VOCAL V	Br. Fernando José Paz Gonzalez
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ LA DEFENSA DE TESIS

DECANA	Inga. Aurelia Anabela Córdoba Estrada
EXAMINADOR	Mtro. Ing. Edgar Darío Álvarez Cotí
EXAMINADOR	Mtro. Ing. Edwin Adalberto Bracamonte Orozco
EXAMINADORA	Mtra. Lcda. Mariela Lizeth Benavides Lázaro
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DETERMINACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE DE MEJOR AJUSTE PARA LA PREDICCIÓN DEL ABSTENCIONISMO ELECTORAL EN GUATEMALA

Tema que me fuera aprobado por la Dirección de la Escuela de Estudios de Postgrado, con fecha 28 de enero de 2020.

Ing. Carlos Ottoniel Guevara Carranza

DTG.658.2021

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Postgrado, al Trabajo de Graduación titulado: **DISEÑO DE MODELO ACTUARIAL PARA LA PROYECCIÓN DE MORTALIDAD EN GUATEMALA**, presentado por el **Ingeniero Carlos Ottoniel Guevara Carranza**, estudiante de la **Maestría en Estadística Aplicada**, y después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:



ing. Anabela Cordova Estrada
Decana



UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA
★

Guatemala, noviembre de 2021.

AACE/cc



Guatemala, noviembre de 2021

LNG.EEP.OI.086.2021

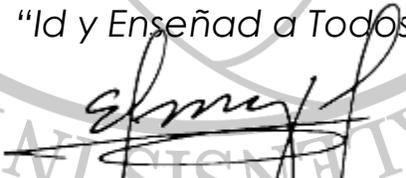
En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

“DETERMINACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE DE MEJOR AJUSTE PARA LA PREDICCIÓN DEL ABSTENCIONISMO ELECTORAL EN GUATEMALA”

presentado por **Carlos Ottoniel Guevara Carranza** quien se identifica con carné **200614153** correspondiente al programa de **Maestría en artes en Estadística aplicada** ; apruebo y autorizo el mismo.

Atentamente,

“Id y Enseñad a Todos”


Mtro. Ing. Edgar Darío Álvarez Cotí
Director



Escuela de Estudios de Postgrado
Facultad de Ingeniería



Guatemala 14 de junio 2021.

M.A. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

M.A. Ingeniero Álvarez Cotí:

Por este medio informo que he revisado y aprobado el Informe Final del trabajo de graduación titulado **“DETERMINACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE DE MEJOR AJUSTE PARA LA PREDICCIÓN DEL ABSTENCIONISMO ELECTORAL EN GUATEMALA”** del estudiante **Carlos Ottoniel Guevara Carranza** quien se identifica con número de carné **2774698800204** del programa de Maestría en Estadística Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014*. Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente,


MSc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestría en Estadística Aplicada
Escuela de Estudios de Postgrado

Guatemala, 16 de noviembre de 2020

Mtro. Ing. Edgar Darío Álvarez Cotí.
Director de la Escuela de Estudios de Postgrado.
Facultad de Ingeniería
Presente.

Distinguido Maestro Álvarez Cotí:

Reciba un cordial saludo deseándole éxitos en sus actividades diarias.

Por medio de la presente hago de su conocimiento que Carlos Ottoniel Guevara Carranza, estudiante de la Maestría en Estadística Aplicada, quien se identifica con carné número 2774698800204, me ha presentado el Informe Final de su trabajo de graduación titulado "**DETERMINACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE DE MEJOR AJUSTE PARA LA PREDICCIÓN DEL ABSTENCIONISMO ELECTORAL EN GUATEMALA**", el cuál realizó bajo mi asesoría.

Luego de revisar el documento que contiene el informe de la investigación, manifiesto que el mismo cumple con los requisitos establecidos, por lo cual le doy mi aprobación.

Sin otro particular, me suscribo a sus respetables órdenes.

Atentamente,



Mayra Virginia Carvajal Castillo
Ingeniera Industrial y
Colegiado No. 13,265

Mayra Virginia Carvajal Castillo
Maestra en Estadística Aplicada

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES.....	III
LISTA DE SÍMBOLOS	VII
GLOSARIO	IX
RESUMEN.....	XI
PLANTEAMIENTO DEL PROBLEMA.....	XIII
OBJETIVOS.....	XVII
RESUMEN DEL MARCO METODOLÓGICO	XIX
INTRODUCCIÓN.....	XXV
1. MARCO REFERENCIAL.....	1
1.1 Estudios Previos	1
2. MARCO TEÓRICO.....	7
2.1 Regresión lineal múltiple.....	7
2.2 Pruebas de normalidad.....	9
2.3 Métodos de correlación.....	100
2.3.1 Coeficiente de correlación de Pearson	11
2.3.2 Coeficiente de correlación de Spearman	12
2.4 Estimadores de bondad de ajuste	13
2.5 Test de Friedman.....	17

2.6	Abstencionismo electoral.....	18
3.	PRESENTACIÓN DE RESULTADOS.....	21
3.1	Análisis descriptivo previo	21
3.2	Análisis de normalidad	32
3.3	Análisis de correlación.....	38
3.4	Análisis de la influencia de las variables que componen el padrón electoral.....	43
3.5	Análisis del mejor ajuste.....	46
4.	DISCUSIÓN DE RESULTADOS	61
4.1	Análisis descriptivo previo	61
4.2	Análisis de normalidad	63
4.3	Análisis de correlación.....	65
4.4	Análisis de la influencia de las variables que componen el padrón electoral.....	67
4.5	Análisis del mejor ajuste.....	69
	CONCLUSIONES.....	73
	RECOMENDACIONES	75
	REFERENCIAS	77
	APÉNDICES.....	83

ÍNDICE DE ILUSTRACIONES

FIGURAS

1.	Distribución de la variable respuesta	22
2.	Número de votos por rango etario vs. alfabetismo	22
3.	Número de votos por rango etario vs. sexo.....	23
4.	Número de votos por cada zona municipal	23
5.	Densidad de la variable respuesta	31
6.	Curva de densidad observada de los residuos del modelo	34
7.	Homocedasticidad y normalidad de los residuos del modelo.....	35
8.	Residuos estudiantizados vs apalancamiento	36
9.	Matriz numérica de la correlación de las c _j variables independientes con la variable respuesta utilizando el coeficiente de Spearman	39
10.	Matriz sombreada de la correlación de las variables independientes con la variable respuesta utilizando el coeficiente de Spearman	40
11.	Matriz de correlación entre las variables del padrón vs variables de la población	41
12.	Correlación de Spearman, IDH vs padrón electoral vs población domiciliar	42

TABLAS

I.	Definición de variables	XX
II.	Descripción de votos I	25
III.	Descripción de votos II	26
IV.	Descripción del padrón I	27
V.	Descripción del padrón II	29
VI.	BIC y AIC como estimadores de la distribución de los votos emitidos	31
VII.	Prueba de Weibull	32
VIII.	Prueba de normalidad Shapiro-Wilk	32
IX.	Residuales	33
X.	Modelo de regresión lineal múltiple	33
XI.	Prueba de Breusch-Pagan estudentizada	36
XII.	Prueba de normalidad de Shapiro-Wilk	36
XIII.	Evaluación de los supuestos del modelo lineal utilizando la prueba global de 4 grados de libertad	37
XIV.	Prueba de suposición de normalidad y de supuesto de homocedasticidad	38
XV.	Prueba de normalidad de Lilliefors (Kolmogorov-Smirnov)	43
XVI.	Prueba de Breusch-Pagan estudentizada	43

XVII.	Evaluación de los supuestos del modelo lineal utilizando la prueba global de 4 grados de libertad.....	44
XVIII.	Suma de rangos del Test de Friedman.....	44
XIX.	Ajuste de empates en el Test de Friedman	45
XX.	Grupos según la probabilidad de diferencias de tratamiento y nivel alfa (0.05)	45
XXI.	Modelos evaluados para todas las personas.....	48
XXII.	Modelos evaluados para personas alfabetas.....	49
XXIII.	Modelos evaluados para personas analfabetas.....	50
XXIV.	Modelos evaluados, para personas de sexo masculino	51
XXV.	Modelos evaluados, para personas de sexo femenino.....	52
XXVI.	Modelos evaluados, para personas entre 18 y 25 años	53
XXVII.	Modelos evaluados, para personas entre 26 y 35 años	54
XXVIII.	Modelos evaluados, para las personas entre 36 y 45 años.....	55
XXIX.	Modelos evaluados, para personas entre 46 y 55 años	56
XXX.	Modelos evaluados, para personas mayores de 55 años.....	57
XXXI.	Resumen de los mejores modelos de cada data analizada.....	58

LISTA DE SÍMBOLOS

Símbolo	Significado
R²	Coeficiente de determinación
Q	Cuartil
sd	Desviación estándar
DAM	Desviación media absoluta
se	Error estándar
RMSE	Error cuadrado medio
AIC	Estadígrafo del criterio de información de Akike
BIC	Estadígrafo del criterio de información de Bayesiano
gl	Grados de libertad
MAE	Media absoluta del error
MR	Media recortada
MIN	Valor mínimo
MAX	Valor máximo

GLOSARIO

Abstencionismo	La no participación en el acto de votar de quienes tienen derecho a ello.
Ajuste	Redefinición de variables y transformación de escalas.
Estado Civil	Relación que tiene la persona física con el estado.
Estimación	Proceso de encontrar una aproximación sobre una medida que se ha de valorar.
Homocedasticidad	Característica que implica que la varianza de los errores es constante a lo largo del tiempo.
IDH	Índice de desarrollo humano.
Padrón electoral	Registro detallado de cada persona que está registrada y tiene derecho a votar.
Rango etario	Clasificación de individuos que están vinculados por la edad.
Regresión	Proceso estadístico para estimar relaciones entre variables.

Zona municipal

Clasificación del Tribunal Supremo Electoral de cada circunscripción municipal dividida en urbana y rural.

RESUMEN

El presente estudio pretende determinar un modelo de regresión lineal múltiple, por medio del análisis de las variables del padrón electoral, para describir el fenómeno del abstencionismo en Guatemala.

Se analizaron las variables del padrón electoral más actualizadas hasta el dos mil veinte así como los resultados de los comicios del dos mil quince en la primera elección, la correlación que tienen estos entre sí y variables socioeconómicas como el IDH.

Para el abstencionismo, se analizó la proporción de votos para calcular un modelo de regresión, agotando los supuestos de linealidad, hasta encontrar el mejor ajustate.

Dentro de los modelos evaluados el modelo de potenciación del gradiente fue el mejor ajustado, dentro de una data en donde las variables como el rango etario, el sexo, el alfabetismo, no se relacionan entre sí, ni con variables socioeconómicas, cuando de abstencionismo se trata en virtud que cada una de ellas se ve afectada por esté fenómeno, sin distinción alguna.

PLANTEAMIENTO DEL PROBLEMA

Contexto general

El padrón electoral está conformado por los ciudadanos inscritos en el Tribunal Supremo Electoral de Guatemala, que están en la capacidad de ejercicio de conformidad a la ley para poder participar en los eventos electorales de la república, sin embargo, en estos eventos electorales, existe un gran número de guatemaltecos que no emiten su voto, por no estar en el padrón, o por no participar en los eventos. Este fenómeno del abstencionismo describe una actividad cívica que afecta todo el territorio nacional, así como las decisiones gubernamentales que afectarían a toda la población durante cuatro años, elegidas por un pequeño sector de la misma.

Descripción del problema

Las campañas electorales difundidas por la institución competente, tienen como propósito incentivar a la población a emitir su sufragio, mediante un evento electoral, sin embargo, aunque la cantidad del padrón electoral crece a un ritmo más acelerado que el de la población, no significa que las tendencias de participación sean altas, por el contrario la apatía a los eventos antes mencionados crece, por lo que es importante definir las proporciones que participan en dicho evento, para efectos de establecer los segmentos de la población que afectan el fenómeno anteriormente relacionado.

En la actualidad no existe ningún modelo estadístico que, por medio de la regresión lineal múltiple, pueda proyectar el fenómeno del abstencionismo electoral en Guatemala, aunque existen varios estudios en otros países, que proyectan dicho fenómeno asociando variables socioeconómicas. Sin embargo dichos modelos no podrían aplicarse a Guatemala, dada la dificultad de poder realizar proyecciones por medio de variables de dicha naturaleza, debido a que estas tienen una varianza tal, que afecta negativamente los estimadores de ajuste, dada la amplia desviación que ostenta la región latinoamericana con respecto de su socioeconomía, por lo que un modelo con base en indicadores socioeconómicos aplicado a Guatemala para proyectar el abstencionismo tendría poca capacidad predictiva.

Formulación del problema

Pregunta Central

¿Qué modelo de regresión lineal múltiple se ajusta mejor al fenómeno del abstencionismo, de conformidad a las variables de los estados civiles de las personas, registradas en el padrón electoral de la República de Guatemala?

Preguntas Auxiliares

- ¿Cuáles son las variables de la población en general, que mayor nivel de asociación tienen con respecto de las variables del padrón electoral?
- ¿Cuáles son los estados civiles que tienen mayor influencia en la explicación de los modelos estadísticos para efectos de proyectar el abstencionismo?

- ¿Qué modelo de regresión lineal múltiple tiene mejor ajuste, con el objeto de explicar de mejor manera el fenómeno del abstencionismo?

Delimitación del problema

El problema se delimitó por los estados civiles tomados en cuenta dentro del padrón electoral para efectos de su caracterización, estableciéndose la tendencia del mismo. Fue necesario conocer el comportamiento de los procesos a nivel nacional, para lo cual se analizaron dos eventos anteriores, dado que los eventos electorales se suscitan cada cuatro años, es menester analizar de dos a tres eventos pasados con el objeto de crear una línea temporal que establezca la tendencia de los segmentos, siempre dependiendo de las necesidades que se fueron dando en el desarrollo de la investigación.

OBJETIVOS

General

Determinar un modelo de regresión lineal múltiple de mejor ajuste, por medio del análisis de las variables del padrón electoral, para describir el fenómeno del abstencionismo en Guatemala

Específicos

- Asociar las variables de la población con respecto al padrón, por medio de la correlación, para identificar la proporción de la población con derecho al voto y su relación entre sí.
- Establecer la influencia de las variables en la proyección del abstencionismo, por medio de los modelos de regresión lineal múltiple, para efectos de la identificación de las variables dominantes en virtud de su capacidad explicativa.
- Determinar el modelo de mejor ajuste, con base en sus indicadores y su expresión parsimónica, para establecer estadísticamente la mayor explicación del fenómeno del abstencionismo,

RESUMEN DEL MARCO METODOLÓGICO

- Características del estudio

El enfoque del estudio propuesto fue cuantitativo, ya que se cuantificó la cantidad de abstención en un futuro, basado en una base de datos comprendida en rangos de edad, cuantificación de votos, número de habitantes, entre otras.

El alcance es correlacional, dado que se buscó determinar la relación que existe entre una variable dependiente y una o más variables independientes. Para identificar esta relación, se postuló una relación funcional entre las variables.

El diseño adoptado fue no experimental, pues la información del padrón electoral conjuntamente con los escrutinios de los eventos electorales se analizó en su estado original sin ninguna manipulación.

- Unidades de análisis

La población en estudio fue el padrón electoral de la república de Guatemala, la cual se encuentra dividida en subpoblaciones descritas en el artículo 5 y 5 bis del Reglamento de la Ley Electoral y de Partidos Políticos, la cual establece la residencia electoral y de conformidad a la ley la circunscripción electoral más pequeña para efectos del cambio de alguna papeleta es la vecindad, por lo que los datos provenientes de los municipios fueron estudiados en su totalidad.

- Variables

Las variables en estudio se describen a continuación:

Tabla I. **Definición de variables**

Variable	Definición teórica	Definición operativa
Padrón Electoral	Documento o registro donde constan el conjunto de ciudadanos a los que la ley reconoce el derecho al sufragio para elegir a sus representantes.	Número de personas inscritas por municipio con derecho a voto.
Votos emitidos	Realización del sufragio, emitida por cualquier persona del padrón electoral.	Número de votos emitidos de conformidad a las resoluciones del TSE.
Población	Total habitantes de la circunscripción territorial	Número de habitantes por la circunscripción municipal.
Edad	Tiempo transcurrido a partir del nacimiento de una persona individual. En este caso, la edad de las personas que tienen derecho a votar.	Frecuencia de las personas que se ubican en los intervalos etáreos establecidos en el padrón electoral.
Alfabetismo	Capacidad para leer y escribir.	Número de alfabetos y analfabetos de las personas del padrón electoral.

Continuación de tabla I.

Sexo	Condición orgánica que distingue a los hombres de las mujeres.	Número de hombres y mujeres de conformidad al padrón electoral.
Residencia electoral	Dirección que se registra para efectos de inclusión en el padrón electoral	Número de municipios dentro de la república de Guatemala.
IDH	Puntuación porcentual de la sumatoria de un tercio de los indicadores de salud, escolaridad y riqueza, para efectos de determinar el desarrollo humano, elaborado por el Programa de las Naciones Unidas para el Desarrollo.	Número que indica el desarrollo humano de cada municipio de conformidad a los criterios de formulación del mismo.

Fuente: elaboración propia.

- Fases del estudio

Fase 1: revisión de literatura

Para la revisión bibliográfica se consultaron periódicamente fuentes teóricas durante todo el proceso de elaboración del estudio, debido a que su contenido estableció una guía, técnica, o precedente orientando dicho evento, dejando constancia de las fuentes, referenciándolas de conformidad a las normas APA.

Fase 2: gestión o recolección de la información

Fue gestionada la información, a través del acceso a la información pública, solicitando los datos contenidos en el padrón electoral al Tribunal Supremo Electoral, durante la primera semana de estudio.

Durante las semanas subsiguientes se dio seguimiento a la solicitud planteada hasta cumplir el plazo que confiere la ley, obteniendo así la información en mención. Cabe destacar que la información obtenida tuvo que ser sometida a un ordenamiento a efecto de poder ser analizada en virtud que la misma expresaba datos que durante el análisis podrían generar una duplicidad.

Fase 3: análisis de información

Comprendió todo el análisis y el ordenamiento de los datos obtenidos para efectos de aplicar las técnicas estadísticas, la finalidad fue en la apreciación de los datos obtenidos para efectos de su ordenamiento dentro de los parámetros que la metodología establece, correlacionando las variables independientes y el abstencionismo, y así determinar el grado de asociación que tienen estas con el fenómeno bajo estudio, proceso que fue programado en la semana inmediata luego de recibir la información, siendo ésta la contenida dentro del padrón electoral así como el IDH y la población total por cada municipio.

Fase 4: interpretación de información

En esta fase se realizaron las pruebas estadísticas de los datos ya ordenados para el estudio.

Para la primera prueba se analizó la normalidad de los datos a través de las pruebas de normalidad Shapiro-Wilks, Kolmogorov-Smirnov Lilliefors y Anderson-Darling, para efectos de correlacionar los datos de conformidad al comportamiento paramétrico (coeficiente de correlación de Pearson) o no paramétrico (coeficiente de correlación de Spearman) en un periodo de tres semanas.

Fase 5: ajuste del modelo

Luego de realizada la inferencia, se analizaron las variables que presentaron una posible explicación a través de una variable independiente, en búsqueda del principio de parsimonia, el cual explica de mejor manera el modelo con una inferencia más simple, es decir, tomando en cuenta menos variables, eliminando las variables que podrían hacer ruido en la modelización, siendo evaluado en la mejora de los estimadores de bondad de ajuste RSME, MAE y R^2 .

Fase 6: elaboración de informe

En esta fase se ejecutaron todos los pasos, que de conformidad a los requerimientos y criterios de forma que la Escuela de Estudio de Postgrado establece a efecto de poder presentar el informe de manera satisfactoria.

INTRODUCCIÓN

El informe final se realizó, analizando el fenómeno del abstencionismo en Guatemala, a través de la data proporcionada por el Tribunal Supremo Electoral, la cual contiene los votos emitidos durante la primera elección del año dos mil quince, y el padrón actualizado al año dos mil diecinueve, con la pretensión de encontrar un modelo de regresión que permita inferir dicho fenómeno de la manera mejor ajustada posible, para tal efecto se modeló se analizaron los estados civiles y el alfabetismo de conformidad a la clasificación del padrón, estableciendo el nivel de asociación de cada uno de los componentes, así como el ajuste del mejor modelo de regresión que pudiera describir de mejor manera el abstencionismo electoral.

Se estructuró en cuatro capítulos, siendo el primero el que referencia la forma de abordar el fenómeno bajo estudio en investigaciones realizadas, sucesivamente el segundo abarca todas las conceptualizaciones teóricas que se ven inmersas dentro de dicho informe tales como los análisis de normalidad, correlación, análisis de regresión, estimadores de ajuste, el análisis de varianza de Friedman y la definición del fenómeno como tal, posteriormente se presentan los resultados de los análisis realizados, finalmente, en el cuarto capítulo se interpretan los resultados obtenidos.

1. MARCO REFERENCIAL

1.1. Estudios Previos

El abstencionismo electoral ha sido descrito a lo largo de la historia dependiendo desde el punto de vista que se le analice. A continuación, se hace mención de diversas investigaciones realizadas que tienen como objeto de estudio las condiciones, variables y características poblacionales que puedan ser vinculantes en la explicación de dicho fenómeno electoral.

Según un estudio sobre el comportamiento electoral en Cataluña en la Revista Española de Investigaciones Sociológicas Riba (2000) afirma:

La aproximación metodológica de la elección racional permite construir grupos de electores que tienen distintas percepciones del espacio en el que se dirime la competición política, hecho que les induce a un comportamiento electoral diferenciado. Combinando estos resultados con los obtenidos del análisis con datos socioestructurales agregados, se establece un cierto perfil de los votantes duales y de los abstencionistas diferenciales. (p. 59)

Estableciendo que existen grupos electorales que marcan cierto comportamiento diferenciado que se refleja a través de un perfil de votantes, es decir, el análisis de los perfiles para este caso del padrón electoral establecen un perfil tal que es posible predecir un comportamiento diferenciado en un evento electoral.

Los fenómenos electorales marginales que describen el abstencionismo y la votación nula en México, pueden ser descritos por medio del análisis

cuantitativo y esbozar las posibles causas que tipifiquen la conducta del abstencionista. (Lutz, 2005). Una de las grandes causas que podrían establecer la variación de este fenómeno es el Índice de Desarrollo Humano (IDH), según un estudio de Fundación Dialnet existe un grado de asociación entre el (IDH) y el abstencionismo, de conformidad a lo expresado por Chavarría (2014) donde comprueba por medio de la correlación entre estas dos variables a nivel cantón como unidad geográfica de análisis, tanto como para las elecciones de alcalde y presidente, incluyendo los mecanismos locales y nacional.

De acuerdo a lo establecido por Instituto Federal Electoral de México [IFE], (1998), conforme a un estudio de la abstención y su correlación con las principales variables sociodemográficas en los eventos electorales de 1994 y 1997, pudieron comprobar los efectos de la competencia en el abstencionismo de la población, mediante la correlación entre los principales caracteres poblacionales y el nivel de participación en los eventos electorales, entre éstas cabe mencionar la escolaridad, ruralidad, ocupación, ingreso y vivienda, las cuales fueron correlacionadas con la participación, con los votos favorables a la oposición comparados con los del gobierno de turno. Aunque ciertas investigaciones apuntan a la vinculación de las variables económicas y sociales para predecir el abstencionismo en América Latina, las variables al correlacionarse con dicho fenómeno hacen poco predecible el modelo según lo establece (Soto, 2016), en el estudio sobre el abstencionismo en México.

Los estudios e investigaciones anteriores sugieren la predicción del fenómeno del abstencionismo, basados en variables socioeconómicas, es decir, las condiciones demográficas e indicadores socioeconómicos podrían ser las variables independientes en los modelos de regresión, dado que en las condiciones que cada una de ellas fue realizada, con ese fin, basándose en las condiciones antes mencionadas y su correlación con el fenómeno de estudio.

Según Hernández (1991), es posible explicar la variabilidad del abstencionismo, conjuntamente con las regiones de mayor participación con un análisis de regresión y correlación ecológica usando variables censales, las cuales fueron analizadas para describir la evolución del padrón electoral de Costa Rica. De conformidad a lo anteriormente expuesto, Lago (2000) en el estudio realizado sobre la abstención diferencial, utiliza modelos de regresión logística, para la determinación de las variables categóricas, tratando de esta manera las variables dependientes multicategóricas u ordinales como una regresión politómica, sugiriendo una modelización lineal múltiple para efectos de su análisis.

De acuerdo con Pons (2019), se puede establecer una relación entre la participación electoral y los porcentajes de voto, a partir del análisis de modelos de regresión lineal y gráficos de dispersión bivariantes. En algunos casos, los estudios que analizan las desigualdades entre las personas que participaron con respecto de las que se abstuvieron, en términos de las causas que actúan en los diferentes plazos del voto, estiman el fenómeno a través de un análisis multivariado; según Santana, Rama, Arellano y Montero (2018), quienes examinaron los principales caracteres sociales y demográficos del perfil de las personas con derecho a voto, entre otras, recogiendo datos sobre las disimilitud entre los participantes y los que se abstuvieron en términos de los factores que predominan en el voto a corto y largo plazo, y estimando un análisis lineal múltiple para proyectar el sufragio a cada una de las coaliciones.

El análisis de regresión lineal múltiple ha sido una herramienta estadística recurrente para el estudio del abstencionismo, según manifiesta Soto (2013), dónde dicho modelo tendría una capacidad explicativa de la ausencia de participación electoral y la ausencia de decisión electoral, comprobando, la

hipótesis de su estudio, la que busca determinar si existen mejoras institucionales de la democracia procedimental a efecto de una apertura del sistema electoral para escoger autoridades legislativas, en el cual pudo establecer la causalidad del fenómeno en mención por medio del método de la regresión lineal múltiple. En ese mismo orden de ideas, un estudio sobre la abstención municipal en Chile para los comicios del 2012 afirma, que las variables más significativas en la participación electoral pueden determinarse a través de la correlación entre las mismas, y mediante el modelo de regresión lineal, (Ramírez, 2013)

Así mismo estos autores, utilizan en sus investigaciones distintos modelos de regresión, cabe destacar que cada una de las peculiaridades de las investigaciones anteriormente citadas, tienen características y tiempos distintos, sin embargo, existe un consenso en la utilización de modelos de regresión lineal múltiple, dado que de esta forma es como se alcanza a explicar mucho mejor el fenómeno, sugiriendo esta como una vía de la investigación realizada.

Cabe destacar que aunque se han realizado varios estudios sobre el abstencionismo y su causalidad, según afirma Fornos, Power y Garand (2004), no encuentran ningún coeficiente de las variables socioeconómicas como estadísticamente significativo derivado que el entorno socioeconómico tiene un efecto más débil en la participación electoral en América Latina que en las democracias industriales avanzadas, lo que sugiere el análisis del estado civil de las personas como indicador de la causalidad del abstencionismo.

De conformidad a lo anteriormente expuesto, existe una tendencia hacia los modelos de regresión lineal multivariada, como la herramienta más utilizada para este tipo de investigaciones, modeladas a partir de variables socioeconómicas, sin embargo estas variables en los países no industrializados como el caso de Guatemala, podrían tener una varianza mucho mayor que aumentaría el error en

la predicción, ya que se estaría analizando sobre una población heterogénea, que aumenta los intervalos de las variables antes mencionadas, por lo que se podría realizar un modelo con respecto de las variables que establece el padrón electoral.

2. MARCO TEÓRICO

2.1. Regresión lineal múltiple

De acuerdo con Walpole, Myers, Myers y Ye (2012), los modelos predictores de regresión, en estadística la regresión lineal o ajuste lineal es una función matemática usada con el fin de aproximar la relación de supeditación entre una variable con dependencia, y las variables sin dependencia, utilizando un sistema de coordenadas cartesianas para reproducir una distribución bidimensional, se obtendrá un cúmulo de puntos que se le denomina comúnmente como diagrama de dispersión, análisis que permitirá estudiar cualitativamente, el nivel de relación entre las dos variables, el paso siguiente, será en la formulación matemática de la dependencia funcional entre ambas variables tanto las dependientes como las que no, que ajusta de mejor manera a la distribución bidimensional. Se llama regresión lineal a la función de carácter lineal, es decir, necesita la definición de dos parámetros: la pendiente y el intercepto en el origen de la recta de regresión, $y=ax+b$.

Como se describe en la cita anterior, en varios empleos existirá más de un regresor, por lo que, más de una variable independiente que ayudará a explicar a Y. Por lo que, si se tratara de explicar las razones de un fenómeno en particular, se esperaría que existan dos variables que puedan determinar el comportamiento de Y, en cuyo caso la estructura múltiple de la regresión se podría escribir como:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 \quad (\text{Ec.1})$$

dónde y es la variable dependiente β_0 es el intercepto, β_1 es el coeficiente de la variable independiente x_1 , β_2 es el coeficiente de la variable independiente x_2 .

Los fenómenos bajo estos supuestos con regresores múltiples, sometidos bajo estudio tienen un análisis resultante que se denomina regresión múltiple; en tanto que el análisis del caso con un solo regresor recibe el nombre de regresión simple.

El coeficiente de determinación, se define como la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretender explicar. Es importante saber que el resultado del coeficiente de determinación oscila entre cero y uno. Cuanto más cerca de uno se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar. De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será. (López, 2017, p.1)

En un modelo lineal general la percepción del coeficiente de determinación se despliega para dar origen a lo que también se denomina como coeficiente de correlación múltiple, este se interpreta como la porción de la variación global de las observaciones comparado con la media, asignable a la función que determina la inferencia estimada. En otras palabras, el R cuadrado es una media del cambio de las variables de predicción incluidas en el modelo, explicando así la variación de las observaciones. Igual que en el modelo lineal simple entre más lejano de cero este el valor, mayor es la variabilidad total que puede explicarse a través de los términos que aparecen en el modelo, por sí mismo el R cuadrado no puede validar el modelo propuesto, según afirma (Canavos, 1988).

2.2. Pruebas de normalidad

Con base en las referencias citadas por Alonso y Montenegro (2015), (Shapiro, Wilk y Chen, Kolmogorov-Smirnov, Kolmogorov, Cramér-von Mises, Cramér, Thadewald y Büning, Jarque y Bera, Razali y Wah, Lilliefors), el desarrollo de las pruebas de normalidad tiene una amplia documentación. A pesar de esto, una gran parte de los estudios bajo diferentes evaluaciones que se realizaron bajo los supuestos de independencia, y son homocedásticos, es decir mantienen la misma varianza entre sí muestran que la prueba de Shapiro-Wilk se desempeña mucho mejor que la prueba de Kolmogorov-Smirnov, por mencionar algunas de las pruebas, dado que cada una tiene distinta potencia con base en las características de la data misma, como el numero de observaciones el supuesto de esféricidad, distribuciones simétricas entre otras.

Cabe destacar que en la cita anteriormente mencionada fueron realizadas simulaciones de Monte Carlo, con el objeto de comparar el comportamiento de cuatro pruebas de normalidad, para lo cual diseñaron un experimento con distintos supuestos, en donde concluyen que la prueba de Shapiro-Wilk en ciertos escenarios tiene mayor potencias que sus pares como Kolmogorov-Smirnov, y Anderson-Darling. Derivado de lo anteriormente expuesto se mencionan los estadígrafos de tres pruebas de normalidad.

De acuerdo con Mohd y Bee (2011), el estadígrafo de Shapiro-Wilk está dado por.

$$W = \frac{\sum_{i=1}^n (a_i x_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (\text{Ec.2})$$

Donde x es el valor de cada muestra desde $i=1$ hasta n y a_i es el valor ordenado que corresponde a la tabla de covarianzas de Shapiro-Wilk.

Estadígrafo de Kolmogorov-Smirnov con los valores críticos de Lilliefors, es

$$D_{n,\alpha} = \frac{\max|F_e(x) - F_o(x)|}{f(n)} \quad (\text{Ec.3})$$

donde F_e es la frecuencia esperada, F_o es la frecuencia observada y

$$f(n) = \frac{0.83+n}{\sqrt{n}} \quad (\text{Ec.4})$$

siendo n el número de muestras.

El estadígrafo de Anderson-Darling se determina por medio de:

$$A^2 = -N - S \quad (\text{Ec.5})$$

Donde N es el tamaño de la muestra y S está dado por:

$$S = \sum_{k=1}^n \frac{2k-1}{N} [\ln F(Y_k) + \ln(1 - F(Y_{N+1-k}))] \quad (\text{Ec.6})$$

F es el valor acumulado de la distribución normal.

2.3. Métodos de correlación

Los métodos de correlación de Spearman y Pearson, que a través de la utilización del álgebra lineal, se admiten como técnicas en un área multivariada con el objeto de establecer similitudes o diferencias entre las variables y los sujetos de evaluación. En otras palabras, establecen un nivel de asociación de

las variables del fenómeno observado contrastadas entre sí. (Restrepo y González, 2007).

2.3.1. Coeficiente de correlación de Pearson

Según Restrepo y González (2007), “el coeficiente de correlación de Pearson tiene como objeto medir el nivel de asociación entre dos variables aleatorias cuantitativas que poseen una distribución normal bivariada conjunta”. (p.185) “Cuando $\rho > 0$ la relación es directa entre las variables. Si $\rho < 0$ la relación es inversa y si $\rho = 0$ son independientes”. (p.185) De esa manera el coeficiente se expresa de cómo detalla su estadígrafo:

$$\rho = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{[\sum_{i=1}^n (X_i - \bar{X})^2][\sum_{i=1}^n (Y_i - \bar{Y})^2]}} \quad (\text{Ec.7})$$

$$-1 \leq \rho \leq 1$$

“El coeficiente de correlación de Pearson es la media geométrica entre las pendientes de los modelos de regresión lineal” (Restrepo y González, 2007, p.185), (Ec.1) se deben cumplir con los supuestos siguientes:

- La distribución combinada de las variables (X, Y) deberá ser normal bivariada
- Tiene que existir una relación lineal dentro de las variables (X, Y) .
- Para cada valor de X , existe un valor de Y distribuidas normalmente.
- Las poblaciones subordinadas con valores en X y Y deben tener una varianza constante.
- Las medias de las subpoblaciones de Y y X tienen una posición en la línea recta.

2.3.2. Coeficiente de correlación de Spearman

El coeficiente de correlación de Spearman es una alternativa no paramétrica al coeficiente de Pearson cuando este no tiene las condiciones suficientes para el cumplimiento de un comportamiento normal. (Restrepo y González, 2007).

$$\rho_S = 1 - 6 \frac{\sum_{i=1}^n d_i^2}{n^3 - n} \quad (\text{Ec.8})$$

$$-1 \leq \rho_S \leq 1$$

Donde:

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n [R_i(x) - R_i(y)]^2 \quad (\text{Ec.9})$$

$$\sum R_i^2(x) = \sum R_i^2(y) = \frac{\eta(\eta+1)(2\eta+1)}{6} \quad (\text{Ec.10})$$

$$\sum_{i=1}^n [R_i(x) - \bar{R}(x)]^2 = \frac{\eta^3 - \eta}{12} \quad (\text{Ec.11})$$

d_i es la diferencia de rangos entre las variables, $R_i(x)$ es el rango de x , $R_i(y)$ es el rango de y , n es el número de la muestra.

Esto refleja una técnica de libre distribución probabilística no paramétrica con una reducción en los supuestos, al hacerlos menos estrictos. La naturaleza misma de esta permite acomodar de un patrón normal, manifestando un nivel de asociación natural entre las variables. (Restrepo y González 2007)

2.4. Estimadores de bondad de ajuste

La bondad de ajuste y los elementos que forman parte para la seleccionar un modelo siempre debe llevar ciertas consideraciones para lo cual Ruiz, Prado y San Martín (2010) afirman, que las funciones para el cálculo de la bondad de ajuste se degradan velozmente conforme al crecimiento del tamaño muestral y la mayoría de investigadores trabajan con muestras reducidas para no degradar los valores de ajuste. Dichos modelos aceptan pocas variables (10-20). Dado el crecimiento en el número de variables, se dificulta calcular adecuadamente las covarianzas observadas.

Como se describe en la cita anterior, cuando las variables aumentan en su número en la misma medida debe crecer el tamaño muestral (es menester una tasa superior a diez individuos por cada variable observada). Varios artículos científicos manifiestan que estos modelos utilizados exceden la capacidad de ajuste y reajuste de las probables asociaciones teóricas, excluyendo e incluyendo las consecuencias y variables de forma tentativa. Por lo que usan los de los parámetros individuales que advierten de las fallas de ajuste que existen en los datos. Estos modelos manipulados en exceso tienden a tener inestabilidad y dejan la bondad de ajuste al momento que se repiten con otros eventos muestrales.

La inferencia estadística en modelos no lineales se basa en las distribuciones asintóticas de las estimaciones de máxima verosimilitud. En general, el estimador de máxima verosimilitud (EMV), para modelos no lineales no es normalmente distribuido, no es insesgado y no es el estimador de mínima varianza. La función que estima la máxima verosimilitud en modelos no lineales tiene propiedades desconocidas para muestras menores. A pesar de esto a medida que se presenta un aumento en el tamaño de la muestra se pueden

predecir las propiedades que el estimador va adquiriendo. Bajo estas circunstancias, la teoría asintótica afirma que la función que calcula la de máxima verosimilitud tiende a ser insesgado, distribuidas normalmente, y se aproxima a la varianza mínima. (Kendall y Stuart 1967).

Luego de modelar un fenómeno por medio de sus variables regresoras, se utilizan varios criterios de decisión con fines de evaluar de un modelo en específico, uno de estos criterios es el criterio de información de Akaike (AIC).

El valor más pequeño de AIC, es el que se estima "más cercano" a la realidad desconocida que generó los datos, entre los modelos candidatos considerados, este parece un concepto muy natural y simple; selecciona la aproximación ajustada al modelo que se estima, en promedio, más cercano al desconocido. Por supuesto, los modelos que contienen valores perdidos quedan fuera de consideración. AIC es útil para seleccionar el mejor modelo del conjunto; sin embargo, si todos los modelos son muy pobres, AIC seguirá seleccionando el que se estima mejor, pero incluso eso podría ser pobre en un sentido absoluto. Por lo tanto, cada esfuerzo debe hacerse para garantizar que el conjunto de modelos esté bien fundado. (Burnham y Anderson 2002).

Estadígrafo del Criterio de Información de Akaike (AIC):

$$AIC = 2k - \ln(L) \quad (\text{Ec.12})$$

donde k es el número de parámetros en el modelo estadístico, y L es el máximo valor de la función de verosimilitud para el modelo estimado

Estadígrafo del Criterio de Información de Bayesiano (BIC):

$$BIC = 2 \ln(L) + k \ln(n) \text{ (Ec.13)}$$

Donde k es el número de parámetros en el modelo estadístico, n es el número de observaciones y L es el máximo valor de la función de verosimilitud para el modelo estimado

Los fundamentos de la parsimonia tienen aplicaciones de importancia en el análisis exploratorio de modelos de regresión lineal múltiple. Dado un grupo de variables que explican para efectos de integrarse a la participación en un modelo bajo estudio, es menester la selección combinada más reducida y simple posible, en virtud de los residuos de la varianza, y la capacidad de predicción son determinantes una posible multicolinealidad.

De acuerdo con Del Valle y Guerra (2012), las técnicas que existen para identificar y considerar la multicolinealidad no consiguen el éxito todas las veces, obligándose que requerir a más de una de éstas, por lo tanto, no se puede invocar una forma única para resolver el problema dado, sin embargo, se pueden sugerir los siguientes pasos para eludir su indeseable efecto:

- Análisis de la matriz de correlaciones entre las variables que explican el modelo, como diagnóstico de una posible multicolinealidad.
- Prueba F parcial y total.
- Error estándar de las funciones que calculan los coeficientes de estimación involucrados.
- La prueba T en los parámetros del modelo.
- Determinación y análisis de los valores y vectores propios.
- Selección de las variables implicadas en la multicolinealidad, mediante el aporte proporcional de los componentes sobre el FIV.

Estadígrafo del Factor de Inflación de la Varianza:

$$FIV_i = \frac{1}{(1-R_i^2)} \text{ (Ec.14)}$$

Donde R_i^2 es el coeficiente de determinación de la regresión de los mínimos cuadrados para cada variable i .

El error cuadrático medio y la media del error absoluto, RMSE y MAE por sus siglas en inglés respectivamente, se utilizan con regularidad en estudios a efecto de estimar los modelos de predicción, algunos trabajos han sugerido que el RMSE, no es un buen indicador del rendimiento promedio del modelo y podría ser engañoso, por lo que sugieren que el MAE puede ser un mejor estimador, mientras que otros estudios establecen lo contrario sugieren que prevalezca el RMSE sobre el MAE, por lo que a menudo se requiere una combinación de ambos estimadores, que incluyan, entre otras, RMSE y MAE para estimar de mejor manera el rendimiento de un modelo.(Chai y Draxler 2014).

$$MAE = \frac{1}{n} \sum_{i=1}^n |e_i| \text{ (Ec.15)}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2} \text{ (Ec.16)}$$

donde n es el número total de la muestra, e es el error de cada muestra desde cada muestra $i=1$ hasta n .

La multicolinealidad no afecta la bondad de ajuste ni la bondad de predicción. En cierto sentido, las variables colineales contienen la misma información sobre la variable dependiente. Si las medidas nominalmente "diferentes" realmente cuantifican el mismo fenómeno, entonces son redundantes. Alternativamente, si las variables reciben diferentes nombres y tal

vez emplean diferentes escalas de medición numérica, pero están altamente correlacionadas entre sí, entonces sufren redundancia. (Belsley 1991).

Se establece que un modelo con capacidad predictiva ostenta homocedasticidad, cuando el error que condiciona a la varianza de las variables que explican el fenómeno, tiene una constancia a lo largo de las observaciones. Wooldridge (2010). Estos criterios anteriormente descritos ajustarían los modelos de regresión para efectos de la determinación de cuál sería el modelo con la mayor capacidad de explicar o proyectar el comportamiento del abstencionismo en el futuro.

2.5. Test de Friedman

Para evaluar la participación de cada uno de los componentes del padrón electoral se puede apoyar en el análisis de varianza en virtud de que analizan los efectos principales de un fenómeno bajo estudio.

La prueba equivalente a un análisis de varianza no paramétrico, basado en los rangos, es la prueba de Friedman, en la cual se considera una distribución para la evaluación de los efectos principales, los cuales se ordenan por filas y bloques, Nuñez-Colin (2018) refiere que los bloques de esta prueba se consideran como una fuente de variación externa por lo que, estos no se hacen generales sino más bien se clasifican dentro de los bloques.

Estadígrafo del test de Friedman:

$$Fr = \left[\left(\frac{12}{bk(k+1)} \right) \sum_{i=1}^k R_i^2 \right] - 3b(k-1) \text{ (Ec.17)}$$

donde b es el número de bloques, k es el número de tratamientos que se está comparando y R_i^2 es la suma de rangos al cuadrado del tratamiento i.

2.6. Abstencionismo electoral

El abstencionismo electoral se enmarca en la no participación en la emisión del sufragio de las personas que posean derecho a ello. El actuar para este caso hace referencia a la no emisión de voto en un proceso refrendatario o electoral. Es una acción que por su naturaleza no tiene consecuencias jurídicas para la persona que ostenta el derecho, con la salvedad de aquellos ordenamientos normativos en que el sufragio se constituya como un deber jurídico, que puede ser exigible por la ley y no como un deber moral o cívico sin consecuencias jurídicas, (Kluwer, 2019).

De acuerdo con la cita anterior el abstencionismo electoral, contiene un fenómeno más extenso que la denominada apatía participativa, la cual es un indicador de la participación política donde se detalla una proporción de los no votantes con respecto del total del padrón siendo estos últimos los que tienen derecho a voto.

Conforme al artículo 224 de la Constitución Política de la República de Guatemala, que se refiere a la división administrativa, el territorio de la república, se divide para su administración en departamentos y estos en municipios, actualmente para al año dos mil veinte, el territorio nacional cuenta con veinte y dos departamentos y trescientos cuarenta municipios.

Para los efectos correspondientes de la emisión del sufragio de cada uno de los ciudadanos con derecho al mismo de cada uno de los municipios del país el artículo 224 de la Ley Electoral y de Partidos Políticos (2016) se establece “del padrón electoral. Con los ciudadanos residentes en cada municipio que se hayan inscrito en el Registro de Ciudadanos, se elaborará un padrón electoral municipal. Cada padrón electoral municipal se identificará con el código del departamento, del municipio y del núcleo poblacional correspondientes” (p 138)

De conformidad al artículo 237 de la Ley Electoral y de Partidos Políticos en el cual preceptúa del escrutinio. Cerrada la votación, los miembros de la Junta Receptora de Votos procederán a la apertura de las urnas y al escrutinio de votos, comprobando que coinciden con el número de votantes, y en su caso, consignando en el acta cualquier diferencia; luego, se procederá a contar los votos emitidos a favor de cada planilla, los votos que se encuentren en blanco y los votos nulos (Decreto 1-85, 2016, p. 143)

Atendiendo lo anteriormente expuesto, el fenómeno del abstencionismo en Guatemala, es una situación de carácter nacional, sin embargo, dada la división administrativa de la república, concatenado con la Ley Electoral y de Partidos Políticos, se establece que las variables dependientes serían las diferencias a nivel municipal del padrón electoral de cada municipio menos el total de votos emitidos, operación que daría como resultado el total de abstenciones. Para determinar el comportamiento del fenómeno antes mencionado, existen herramientas y modelos estadísticos que pueden, predecir el abstencionismo.

3. PRESENTACIÓN DE RESULTADOS

3.1. Análisis descriptivo previo

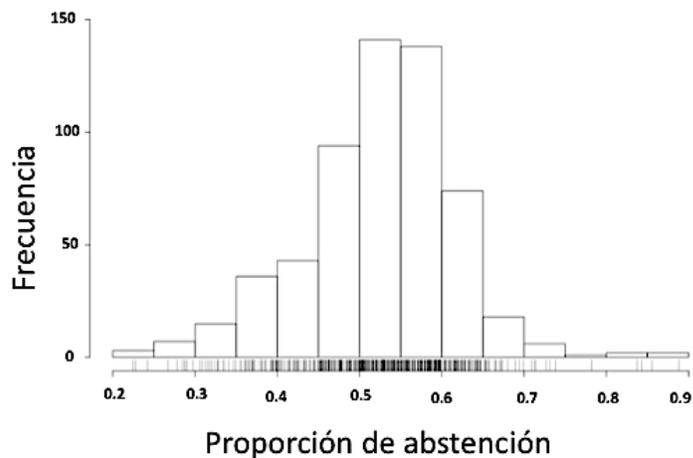
Los datos analizados para este estudio, se resumen en dos datas, las más actualizadas hasta el dos mil veinte cabe destacar que por la naturaleza del comportamiento se exceptúa del estudio las Consultas Populares, y la Segunda Elección derivado que estas no tienen el mismo comportamiento que una Primera Elección.

La primera data consiste en el número de votos emitidos por cada rango etario, para esta clasificación el Tribunal Supremo Electoral usó cinco rangos etarios, divididos todos estos a su vez por sexo y alfabetismo, conteniendo una cantidad de votos emitidos por municipio en sus zonas urbanas y en su zona rural si así la tuviere, haciendo un total de quinientas cuarenta observaciones, basado en el evento electoral disponible con registro actualizado, que corresponde a los comicios del dos mil quince de la República de Guatemala.

La segunda data, al igual que la primera, contiene rangos etarios, sólo que esta contiene once rangos, divididos todos estos a su vez por sexo y alfabetismo, teniendo una observación por cada vecindad para un total de trescientas cuarenta observaciones, por cada variable, a diferencia de la primera ésta última contiene el padrón actualizado hasta el año dos mil diecinueve, por lo que las predicciones de este estudio fueron inferidas a partir de los votos con el objeto de predecir el abstencionismo, basados en la información de la data que corresponde a los votos emitidos del evento electoral del año dos mil quince, y

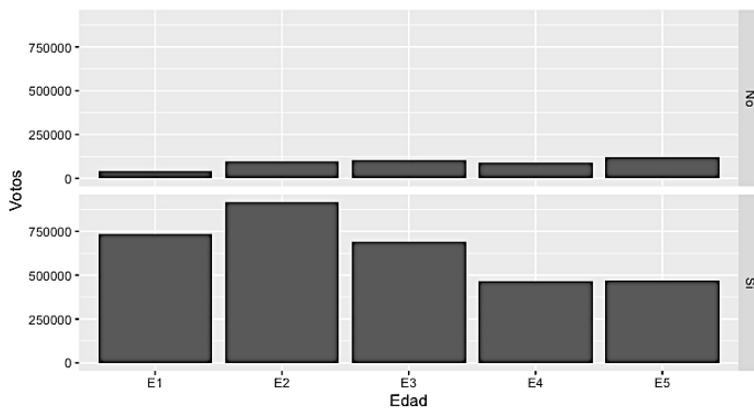
para la composición del padrón electoral se evaluó, la data que contiene el padrón actualizado hasta el año dos mil diecinueve.

Figura 1. **Distribución de los votos emitidos**



Fuente: elaboración propia.

Figura 2. **Número de votos por rango etario vs. alfabetismo**



Fuente: elaboración propia.

Donde:

E1: personas inscritas en el padrón electoral entre (18-25) años

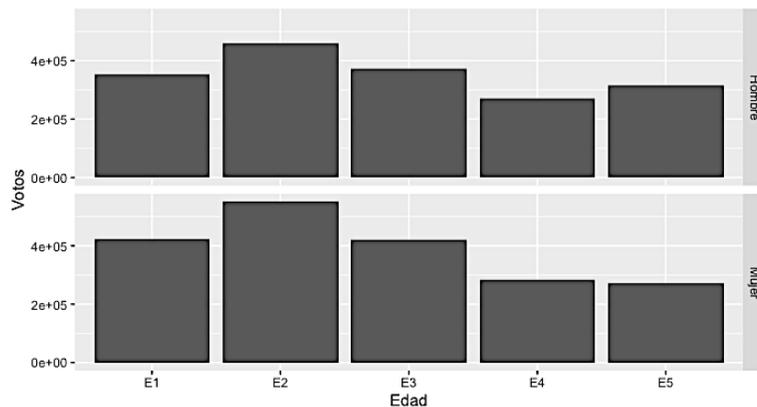
E2: personas inscritas en el padrón electoral entre (26-35) años

E3: personas inscritas en el padrón electoral entre (36-45) años

E4: personas inscritas en el padrón electoral entre (46-55) años

E5: personas inscritas en el padrón electoral mayores a 55 años

Figura 3. **Número de votos por rango etario vs. sexo**



Fuente: elaboración propia.

Donde:

E1: personas inscritas en el padrón electoral entre (18-25) años

E2: personas inscritas en el padrón electoral entre (26-35) años

E3: personas inscritas en el padrón electoral entre (36-45) años

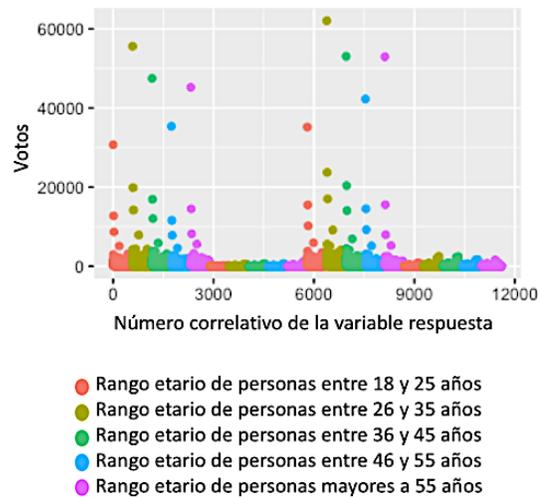
E4: personas inscritas en el padrón electoral entre (46-55) años

E5: personas inscritas en el padrón electoral mayores a 55 años

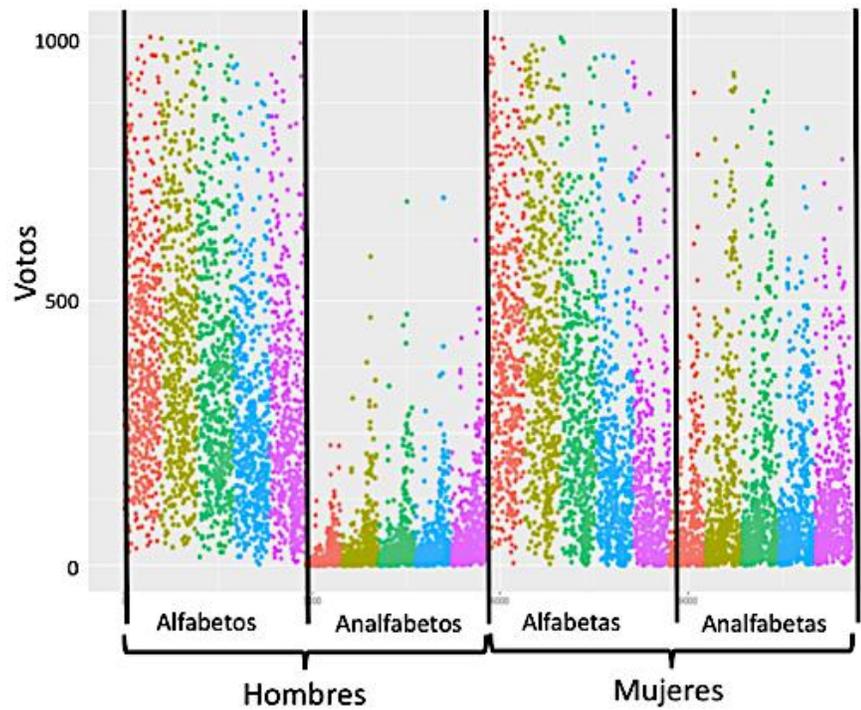
A continuación, se muestra una figura, en la cual se representa el número total de votos emitidos, por cada zona municipal, que comprenden la urbana y la rural si la hubiere, así mismo se segmenta por sexo y alfabetismo, haciendo una ampliación entre el rango de cero a mil votantes ya que es donde se encuentra la mayor cantidad de votos de conformidad a la dispersión antes mencionada.

Figura 4.

Número de votos emitidos por cada zona municipal



Ampliación del rango de 0 a 1000 votos



Fuente: elaboración propia.

A continuación, se muestra un análisis descriptivo de la primera data.

Tabla II. Descripción de votos I

Variable	Media	Mediana	MR	DAM	MIN	MAX
E1HA	594.65	404.5	438.14	254.27	26	30715
E2HA	761.66	451.5	496.32	325.43	34	55576
E3HA	605.92	342.5	379.88	238.7	16	47481
E4HA	430.42	234.5	267.12	176.43	2	35375
E5HA	474.52	246.5	280.03	199.41	6	45205
E1HAN	14.35	5	8.01	5.93	0	227
E2HAN	31.38	11	17.42	13.34	0	584
E3HAN	35.81	13	20.28	14.83	0	688
E4HAN	35.86	16	22.65	17.79	0	695
E5HAN	68.45	38.5	49.92	39.29	0	1182
E1MA	669.32	440.5	475.95	309.12	4	35192
E2MA	815.85	434	500.4	361.01	3	62052
E3MA	581.4	271.5	316.71	252.04	1	53059
E4MA	370.54	144	174.52	146.78	0	42268
E5MA	329.58	91.5	122.89	113.42	0	52966
E1MAN	57.4	16	31.79	19.27	0	894
E2MAN	133.19	45	80.84	53.37	0	2532
E3MAN	142	59	93.32	66.72	0	2493
E4MAN	115.84	60	84.38	63.01	1	1657
E5MAN	136.68	88.5	109.89	81.54	2	1595

Fuente: elaboración propia.

Donde:

H: personas inscritas en el padrón electoral de sexo masculino

M: personas inscritas en el padrón electoral de sexo femenino

A: personas inscritas en el padrón electoral alfabetas

AN: personas inscritas en el padrón electoral analfabetas

E1: personas inscritas en el padrón electoral entre (18-25) años

E2: personas inscritas en el padrón electoral entre (26-35) años

E3: personas inscritas en el padrón electoral entre (36-45) años

E4: personas inscritas en el padrón electoral entre (46-55) años

E5: personas inscritas en el padrón electoral mayores a 55 años

Tabla III. Descripción de votos II

Variable	Sd	Rango	Agudez	Curtosis	Error estándar
E1HA	1461.24	30689	16.47	318.29	60.67
E2HA	2565.12	55542	17.85	363.27	106.51
E3HA	2184.09	47465	18.06	369.57	90.69
E4HA	1604.57	35373	18.61	389.77	66.63
E5HA	2025.02	45199	19.29	411.87	84.08
E1HAN	27.03	227	4.05	20.72	1.12
E2HAN	59.59	584	4.28	24.26	2.47
E3HAN	66.16	688	4.34	26.11	2.75
E4HAN	59.08	695	4.58	33.39	2.45
E5HAN	93.59	1182	4.59	38.25	3.89
E1MA	1703.16	35188	15.9	299.94	70.72
E2MA	2908.55	62049	17.3	344.22	120.77
E3MA	2474.29	53058	17.67	354.57	102.74
E4MA	1922.92	42268	18.73	391.5	79.84
E5MA	2332.81	52966	20.37	447.25	96.86
E1MAN	104.51	894	3.51	16.02	4.34
E2MAN	232.82	2532	4.43	30.53	9.67
E3MAN	227.63	2493	4.55	33.9	9.45
E4MAN	158.32	1656	4.11	29.52	6.57
E5MAN	149.84	1593	3.27	20.06	6.22

Fuente: elaboración propia.

Donde:

H: personas inscritas en el padrón electoral de sexo masculino

M: personas inscritas en el padrón electoral de sexo femenino

A: personas inscritas en el padrón electoral alfabetas

AN: personas inscritas en el padrón electoral analfabetas

E1: personas inscritas en el padrón electoral entre (18-25) años

E2: personas inscritas en el padrón electoral entre (26-35) años

E3: personas inscritas en el padrón electoral entre (36-45) años

E4: personas inscritas en el padrón electoral entre (46-55) años

E5: personas inscritas en el padrón electoral mayores a 55 años

A continuación, se muestra un análisis descriptivo de la segunda data

Tabla IV. Descripción del padrón I

Variable	Media	Desviación Estándar	Mediana	Media Recortada	Desviación Media Absoluta
E1HA	1230.28	2224.4	867	951.66	527.81
E2HA	1342.77	2236.08	935	1038.82	639.74
E3HA	1327.57	2742.41	843	966.76	595.26
E4HA	1156.67	2615.87	697	807.86	495.19
E5HA	982.89	2294.05	594	676.66	412.9
E6HA	804.09	1943.64	480.5	550.56	343.96
E7HA	655.66	1683.07	386	440.62	272.06
E8HA	559.07	1624.01	309.5	363.95	220.91
E9HA	422.98	1273.67	243	270.59	174.21
E10HA	314.2	1002.29	166.5	195.56	118.61
E11HA	458.73	1548.37	248.5	280.68	192
E1HAN	55.09	67.34	31	41.61	28.17
E2HAN	104.3	120.68	60	80.06	56.34
E3HAN	136.95	162.17	79	104.51	73.39
E4HAN	160.03	198.49	92.5	119.11	85.99
E5HAN	161.37	195.97	95	121.48	91.92
E6HAN	143.63	167.59	91	110.77	86.73
E7HAN	148.51	163.43	97	118.04	88.96
E8HAN	161.38	167.32	111.5	131.79	103.78
E9HAN	148.08	157.82	103	121	91.92
E10HAN	142.56	148.97	98	117.44	91.18
E11HAN	322.08	280.29	245.5	279.78	199.41
E1MA	1487	2491.08	1033	1154.09	625.66
E2MA	1562.38	2599.85	1057	1176.4	702.01
E3MA	1443.59	3032.72	885.5	1016.26	630.85
E4MA	1198.47	2944.82	699.5	787.5	489.26
E5MA	956.87	2578.74	528.5	605.14	389.18
E6MA	743.66	2182.83	385	454.06	286.14
E7MA	579.75	1920.13	270.5	331.71	220.17
E8MA	479.43	1888.34	197.5	251.11	179.39

Continuación tabla IV

E9MA	348.41	1531.49	137.5	171.15	126.02
E10MA	248.87	1212.37	91	114.27	90.44
E11MA	358.71	2049.75	113	150.14	124.54
E1MAN	113.56	152.98	49.5	78.76	49.67
E2MAN	229.83	290.79	113.5	168.65	114.9
E3MAN	308.86	384.35	167.5	231.6	169.76
E4MAN	372.76	441.56	214	289.3	202.37
E5MAN	374.06	424.08	230.5	295.99	220.17
E6MAN	336.55	362.17	212.5	272.18	198.67
E7MAN	314.39	318.34	213	257.01	188.29
E8MAN	320.52	311.83	224	266.09	194.22
E9MAN	276.44	268.77	200	229.4	169.02
E10MAN	239.19	233.79	169.5	199.03	138.62
E11MAN	441.76	440.14	333.5	367.13	246.11

Fuente: elaboración propia.

donde:

H: personas inscritas en el padrón electoral de sexo masculino

M: personas inscritas en el padrón electoral de sexo femenino

A: personas inscritas en el padrón electoral alfabetas

AN: personas inscritas en el padrón electoral analfabetas

E1: personas inscritas en el padrón electoral entre (18-25) años

E2: personas inscritas en el padrón electoral entre (26-30) años

E3: personas inscritas en el padrón electoral entre (31-35) años

E4: personas inscritas en el padrón electoral entre (36-40) años

E5: personas inscritas en el padrón electoral entre (41-45) años

E6: personas inscritas en el padrón electoral entre (46-50) años

E7: personas inscritas en el padrón electoral entre (51-55) años

E8: personas inscritas en el padrón electoral entre (56-60) años

E9: personas inscritas en el padrón electoral entre (61-65) años

E10: personas inscritas en el padrón electoral entre (66-70) años

E11: personas inscritas en el padrón electoral mayores a 70 años

Tabla V. Descripción del padrón II

Variable	Mínimo	Máximo	Rango	Coefficiente de Asimetría	Curtosis	se
E1HA	154	35825	35671	12.03	174.21	120.63
E2HA	112	35285	35173	11.18	157.49	121.27
E3HA	81	45271	45190	12.67	193.19	148.73
E4HA	87	43007	42920	12.75	192.92	141.87
E5HA	59	37863	37804	12.91	196.65	124.41
E6HA	51	32555	32504	13.39	208.89	105.41
E7HA	41	28544	28503	13.84	220.6	91.28
E8HA	29	28069	28040	14.65	240.54	88.07
E9HA	25	22047	22022	14.74	242.72	69.07
E10HA	19	17360	17341	14.84	244.53	54.36
E11HA	13	26983	26970	15.12	251.6	83.97
E1HAN	0	594	594	3.3	16.28	3.65
E2HAN	1	826	825	2.63	8.81	6.54
E3HAN	0	1197	1197	2.76	10.15	8.8
E4HAN	1	1607	1606	3.05	13.02	10.76
E5HAN	3	1540	1537	3.03	13.01	10.63
E6HAN	1	1368	1367	2.96	12.83	9.09
E7HAN	1	1402	1401	2.96	13.86	8.86
E8HAN	1	1441	1440	2.75	12.63	9.07
E9HAN	3	1598	1595	3.53	22.88	8.56
E10HAN	0	1471	1471	3.38	20.78	8.08
E11HAN	5	2318	2313	2.24	8.83	15.2
E1MA	177	39120	38943	11.25	156.23	135.1
E2MA	114	39741	39627	10.42	139.48	141
E3MA	104	48905	48801	12.05	177.01	164.47
E4MA	53	47431	47378	12.31	180.38	159.71
E5MA	30	41884	41854	12.65	188.1	139.85
E6MA	19	36062	36043	13.2	202.15	118.38
E7MA	8	32182	32174	13.74	215.74	104.13
E8MA	3	32469	32466	14.73	241.04	102.41
E9MA	2	26581	26579	15.16	251.74	83.06
E10MA	0	21181	21181	15.48	259.79	65.75
E11MA	0	36445	36445	16.26	280.33	111.16

Continuación tabla V

E1MAN	0	845	845	2.35	5.72	8.3
E2MAN	1	1878	1877	2.53	8.04	15.77
E3MAN	0	3023	3023	2.86	11.65	20.84
E4MAN	4	3726	3722	2.94	13.47	23.95
E5MAN	2	3422	3420	2.84	12.36	23
E6MAN	4	2783	2779	2.67	10.82	19.64
E7MAN	9	2305	2296	2.48	8.79	17.26
E8MAN	8	2258	2250	2.57	9.53	16.91
E9MAN	11	2278	2267	2.87	12.81	14.58
E10MAN	8	2298	2290	3.46	20.33	12.68
E11MAN	31	5268	5237	4.92	43.06	23.87

Fuente: elaboración propia.

Donde:

H: personas inscritas en el padrón electoral de sexo masculino

M: personas inscritas en el padrón electoral de sexo femenino

A: personas inscritas en el padrón electoral alfabetas

AN: personas inscritas en el padrón electoral analfabetas

E1: personas inscritas en el padrón electoral entre (18-25) años

E2: personas inscritas en el padrón electoral entre (26-30) años

E3: personas inscritas en el padrón electoral entre (31-35) años

E4: personas inscritas en el padrón electoral entre (36-40) años

E5: personas inscritas en el padrón electoral entre (41-45) años

E6: personas inscritas en el padrón electoral entre (46-50) años

E7: personas inscritas en el padrón electoral entre (51-55) años

E8: personas inscritas en el padrón electoral entre (56-60) años

E9: personas inscritas en el padrón electoral entre (61-65) años

E10: personas inscritas en el padrón electoral entre (66-70) años

E11: personas inscritas en el padrón electoral mayores a 70 años

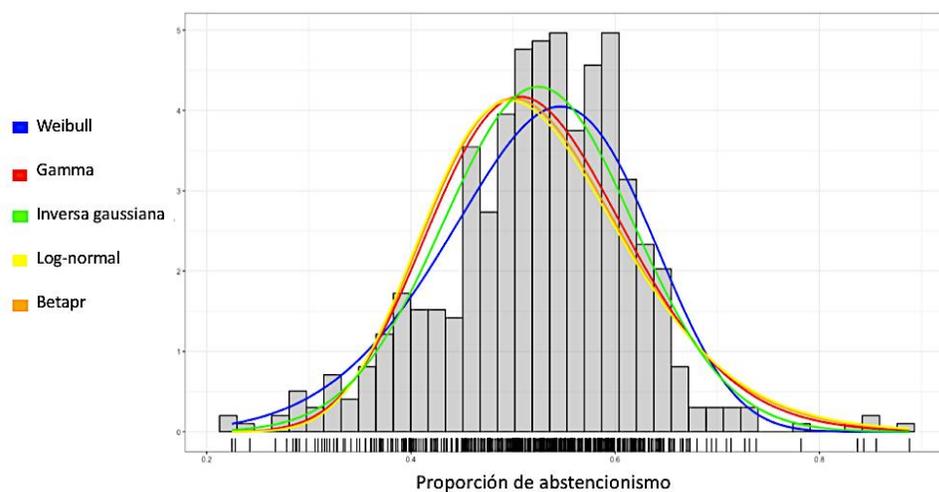
Como primera aproximación a la normalidad de los datos se realizó una prueba sobre la distribución de la variable respuesta, en este caso votos emitidos, estimándola a través de los parámetros del Criterio de Información Bayesiana (BIC) y el Criterio de Información de Akaike (AIC), con el objeto de determinar la distribución que mejor describe la distribución de la misma.

Tabla VI. **BIC y AIC como estimadores de la distribución de los votos emitidos**

	Distribución	gl	BIC	AIC
1	Weibull	2	-1078.6244	-1087.3505
2	Gamma	2	-1060.7629	-1069.4890
3	Betapr	2	-1041.5163	-1050.2424
4	Log-normal	2	-1028.9153	-1037.6413
5	Inversa gaussiana	2	-1024.7506	-1033.4767
6	Gamma inversa	2	-985.5025	-994.2286
7	Weibull inversa	2	-745.4261	-754.1522
8	Rayleigh	1	-330.1748	-334.5379
9	Exponencial	1	418.1278	413.7648

Fuente: elaboración propia.

Figura 5. **Densidad de los votos emitidos**



Fuente: elaboración propia.

Posteriormente, se realizó un test de la mejor distribución en este caso el Test de Weibull derivado que la aproximación de los criterios de densidad de la distribución la cual no necesariamente refleja el comportamiento de la misma, ya que esta distribución es la mejor apegada al comportamiento del fenómeno estudiado. Lo cual se comprueba con el estadígrafo de prueba que a continuación se muestra.

Tabla VII. **Prueba de Weibull**

Data	Variable respuesta
Correlación	0.99544
Valor de p	0.09363

Fuente: elaboración propia.

Derivado que el p valor del test de Weibull es mayor a la tolerancia del cinco por ciento se constata que la distribución de la variable respuesta se ajusta de mejor manera a una distribución de Weibull.

3.2. **Análisis de normalidad**

Como primer paso en el análisis de la normalidad se realizó un test sobre la variable respuesta en este caso votos emitidos, para lo cual se realizó la prueba de normalidad de Shapiro-Wilk.

Tabla VIII. **Prueba de normalidad Shapiro-Wilk**

Data	Variable respuesta
Estadígrafo de Shapiro-Wilk (W)	0.9627772
Valor de p	$5.2150000 \times 10^{-11}$

Fuente: elaboración propia.

Dado que el p valor del test es inferior al nivel de tolerancia del cinco por ciento se establece que la variable respuesta no tiene un comportamiento normal, en virtud de la determinación del cumplimiento de la normalidad de los supuestos, se corrió el modelo de regresión lineal múltiple para los efectos de la determinación antes mencionados. En la que se muestra los residuos del modelo en la Tabla VII y los coeficientes del mismo en la Tabla VIII.

Tabla IX. **Residuales**

Mínimo	1Q	Mediana	3Q	Máximo
-0.28292	0.04224	0.00373	0.04077	0.47646

Fuente: elaboración propia.

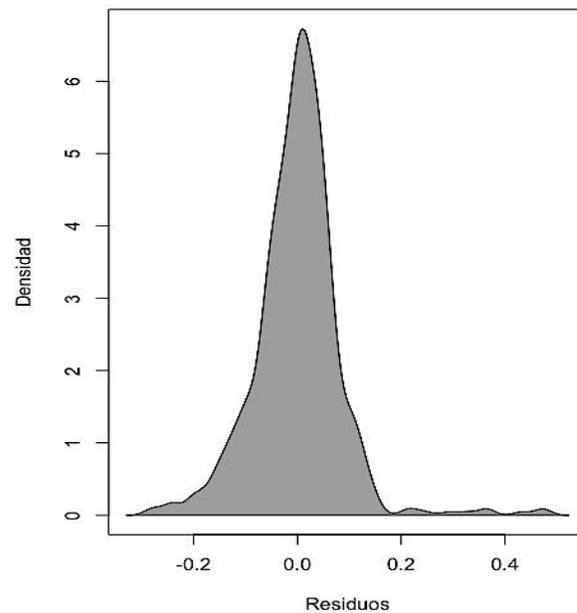
Tabla X. **Modelo de regresión lineal múltiple**

Coeficientes	Estimación	Error estándar	Valor t	Pr(> t)
Intercepto	5.220×10^{-01}	6.810×10^{-03}	76.671	$< 2 \times 10^{-16}$
E1HA	-7.590×10^{-05}	6.450×10^{-05}	-1.177	0.239
E2HA	-3.010×10^{-05}	7.340×10^{-05}	-0.410	0.682
E3HA	7.110×10^{-05}	9.260×10^{-05}	0.768	0.442
E4HA	-5.610×10^{-05}	1.120×10^{-04}	-0.500	0.617
E5HA	6.720×10^{-05}	5.980×10^{-05}	1.124	0.261
E1HAN	5.850×10^{-04}	5.870×10^{-04}	0.997	0.319
E2HAN	-5.730×10^{-05}	5.310×10^{-04}	-0.108	0.914
E3HAN	3.460×10^{-04}	5.470×10^{-04}	0.631	0.527
E4HAN	-9.570×10^{-04}	6.120×10^{-04}	-1.564	0.118
E5HAN	8.180×10^{-04}	2.200×10^{-04}	3.726	0.002
E1MA	-9.930×10^{-05}	6.350×10^{-05}	-1.565	0.118
E2MA	1.630×10^{-05}	7.990×10^{-05}	0.204	0.838
E3MA	1.140×10^{-04}	1.060×10^{-04}	1.079	0.280
E4MA	-1.270×10^{-04}	1.490×10^{-04}	-0.853	0.394
E5MA	2.240×10^{-05}	5.110×10^{-05}	0.439	0.660
E1MAN	-1.950×10^{-04}	2.290×10^{-04}	-0.851	0.395
E2MAN	1.670×10^{-04}	2.070×10^{-04}	0.805	0.421
E3MAN	-7.070×10^{-04}	2.370×10^{-04}	-2.979	0.003
E4MAN	9.210×10^{-04}	2.680×10^{-04}	3.438	0.006
E5MAN	-6.900×10^{-05}	1.210×10^{-04}	-0.572	0.567

Fuente: elaboración propia.

A continuación se muestra la curva de la densidad de los residuos del modelo en la cual se grafica la distribución de las distancias que existen entre la recta proyectada por el modelo y el dato real.

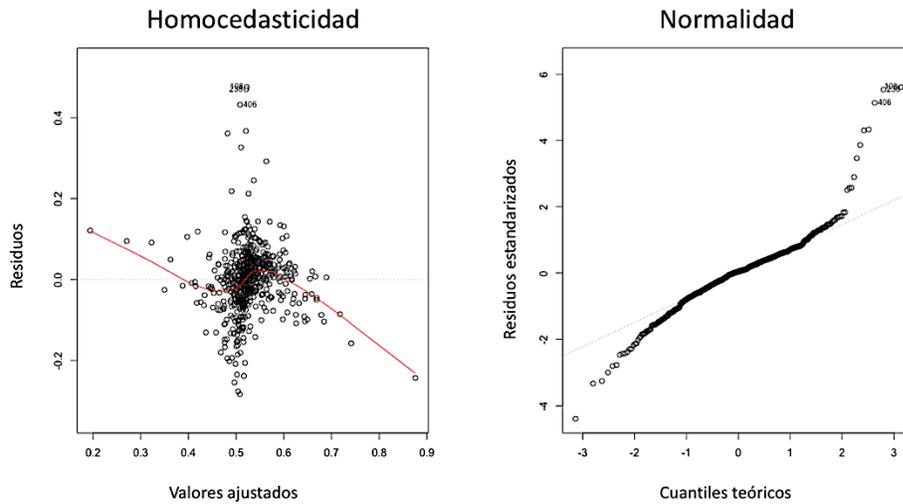
Figura 6. **Curva de densidad observada de los residuos del modelo**



Fuente: elaboración propia.

Como se observa en la figura 6, la distribución de esos residuos debe cumplir con ciertos parámetros, al no cumplirlos se denomina no paramétrica derivado que ausenta las características suficientes para ostentar dicha calidad por lo consiguiente se evalúa su homocedasticidad y normalidad, como lo muestra la figura 7.

Figura 7. **Homocedasticidad y normalidad de los residuos del modelo**

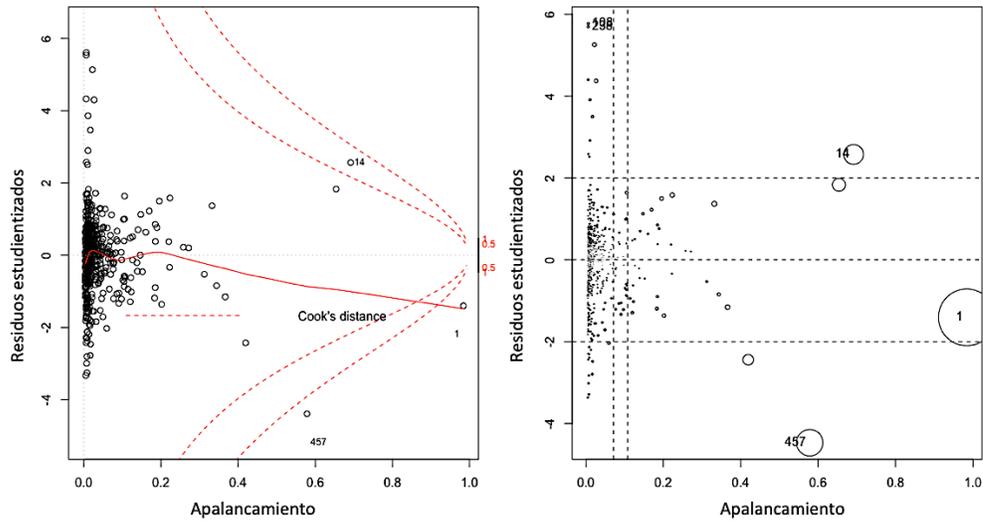


Fuente: elaboración propia.

Los criterios para la determinación de los parámetros están dados por los estadígrafos de prueba, y no por una apreciación gráfica, sin embargo los desfares en la aleatoriedad y en el apego a la recta que se muestran en la anterior figura, hacen alusión a un comportamiento no paramétrico, que como se indica debe ser corroborado a través de un estadígrafo de prueba para el efecto.

En el mismo orden de ideas se anlaizan las observaciones determinantes, para este caso una observación determinante regularmente es una observacion que tiene una distancia mucho más amplias que sus homologas a tal efecto que genera un error mucho mayor, esto claro está es un efecto común derivado que en algunos lugares la cantidad de votantes es muy superior a las demás, como se grafica en la figura que se presenta a continuación.

Figura 8. **Residuos estudentizados vs apalancamiento**



Fuente: elaboración propia.

Posteriormente, se presentan los análisis de los supuestos del modelo representado en la tabla VIII.

Tabla XI. **Prueba de Breusch-Pagan estudentizada**

Data	Residuos del modelo
Estadígrafo de Breusch-Pagan (BP)	37.13000×10^9
Grados de libertad	20.00000
Valor de p	0.01126

Fuente: elaboración propia.

Tabla XII. **Prueba de normalidad de Shapiro-Wilk**

Data	Residuos del modelo
Estadígrafo de Shapiro-Wilk (W)	0.91317
Valor de p	$<2.20000 \times 10^{-16}$

Fuente: elaboración propia.

Tabla XIII. **Evaluación de los supuestos del modelo lineal utilizando la prueba global de 4 grados de libertad**

	Valor del Estadígrafo	Valor de p	Calificación del supuesto
Estadística global	1034.76	0.00	No cumple
Oblicuidad	68.66	1.11×10^{-16}	No cumple
Curtosis	905.30	0.00	No cumple
Función de enlace	12.49	4.10×10^{-4}	No cumple
Heteroscedasticidad	48.31	3.63×10^{-12}	No cumple

Fuente: elaboración propia.

Con base en los análisis anteriormente presentados se afirma que el modelo de regresión lineal múltiple, no cumple con la normalidad requerida, no cumple con la homocedasticidad requerida y tampoco cumple los criterios presentados de la evaluación de la prueba global, cabe destacar que el modelo presenta cinco valores de apalancamiento, el dato uno, catorce, ciento ocho, doscientos treinta y ocho y cuatrocientos cincuenta y siete lo que indica que estas observaciones son influyentes en el modelaje.

Con el objeto de agotar todas las vías de la normalidad se procedió a realizar las transformaciones posibles con el software estadístico a efecto de encontrar una de estas que pudiera adecuarse a la normalidad.

El rango lambda predeterminado para la transformación de desplazamiento de raíz cuadrada se calcula en función del rango de datos. El valor inferior se establece en 0.7736122349611350 y el valor superior en 0.775193798449612

Tabla XIV. **Prueba de suposición de normalidad y de supuesto de homocedasticidad**

Transformación	Agudez	Curtosis	(W)	W p valor	BP	BP p valor
Sin transformar	0.8399	9.0995	0.9132	0.0000	37.1391	0.0113
Boxcox	0.4175	7.8582	0.9245	0.0000	41.8715	0.0029
Dual	-0.5925	7.1128	0.9213	0.0000	44.8979	0.0011
Log	-0.5925	7.1128	0.9213	0.0000	44.8979	0.0011
Bickeldoksum	0.4175	7.8582	0.9245	0.0000	41.8715	0.0029
Gpower	0.2556	7.2014	0.9323	0.0000	44.5368	0.0013
Manly	0.2087	7.1445	0.9324	0.0000	44.8819	0.0011
Modulus	0.2808	7.3945	0.9295	0.0000	43.7872	0.0016
Yeojohnson	0.2808	7.3945	0.9295	0.0000	43.7872	0.0016
Reciprocal	-1.9798	11.6514	0.8460	0.0000	28.3740	0.1008
Neglog	0.3216	7.4888	0.9287	0.0000	43.3420	0.0018
Glog	0.4878	7.7972	0.9267	0.0000	41.7686	0.0030

Fuente: elaboración propia.

Donde:

W: es el estadígrafo de Shapiro-Wilk

BP: es el estadígrafo de Breusch-Pagan

Como lo muestra la tabla que antecede ninguna de las transformaciones posibles pudo cumplir con los supuestos de normalidad y homocedasticidad por lo que se determina que los datos no pueden ser modelados paramétricamente.

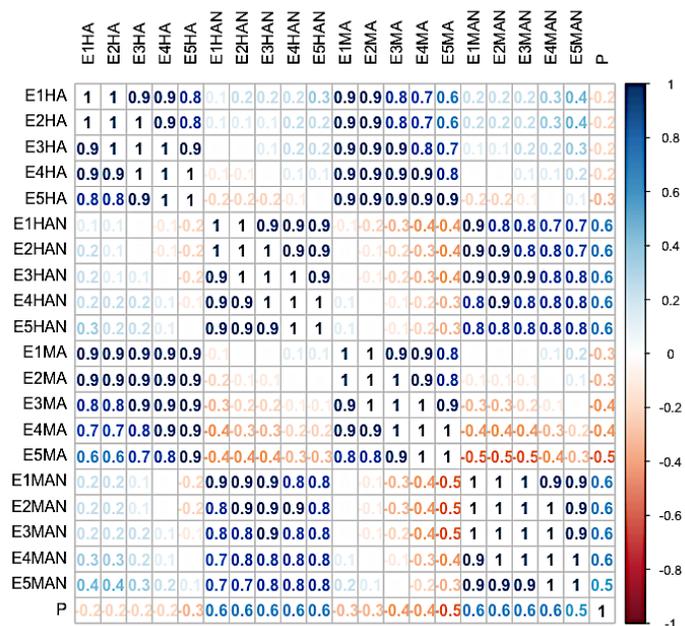
3.3. Análisis de correlación

Basado en que en la sección anterior se determinó que no hay normalidad en los datos se procedió a correlacionar por medio del coeficiente no paramétrico

de Spearman, cada una de las variables del padrón electoral que contiene el conteo de votos y con la proporción de votantes..

A continuación se muestra en la Figura 9 y 10 la correlación de las variables independientes y la variable de respuesta, proporción de votos emitidos, utilizando el coeficiente de Spearman.

Figura 9. **Matriz numérica de la correlación entre las variables**



Fuente: elaboración propia.

Donde:

H: personas inscritas en el padrón electoral de sexo masculino

M: personas inscritas en el padrón electoral de sexo femenino

A: personas inscritas en el padrón electoral alfabetas

AN: personas inscritas en el padrón electoral analfabetas

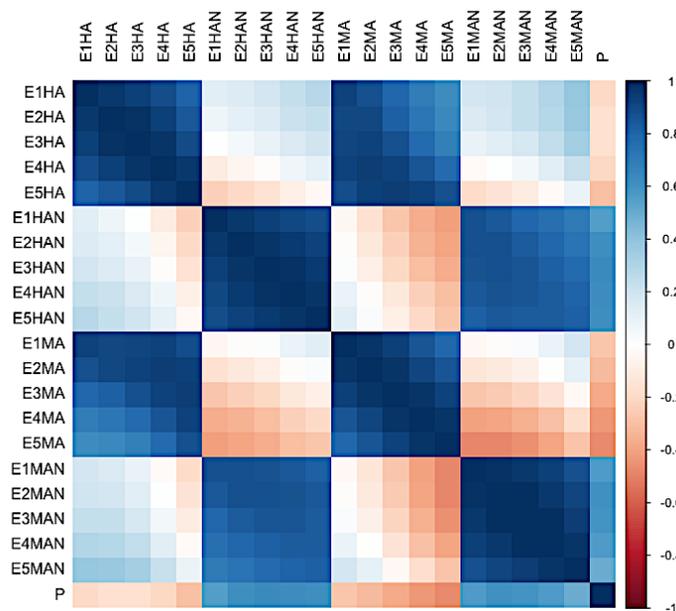
E1: personas inscritas en el padrón electoral entre (18-25) años

E2: personas inscritas en el padrón electoral entre (26-35) años

E3: personas inscritas en el padrón electoral entre (36-45) años

E4: personas inscritas en el padrón electoral entre (46-55) años
 E5: personas inscritas en el padrón electoral mayores a 55 años
 P: es la proporción de abstinentes (variable respuesta)

Figura 10. **Matriz sombreada de la correlación entre las variables**



Fuente: elaboración propia.

Donde:

H: personas inscritas en el padrón electoral de sexo masculino

M: personas inscritas en el padrón electoral de sexo femenino

A: personas inscritas en el padrón electoral alfabetas

AN: personas inscritas en el padrón electoral analfabetas

E1: personas inscritas en el padrón electoral entre (18-25) años

E2: personas inscritas en el padrón electoral entre (26-35) años

E3: personas inscritas en el padrón electoral entre (36-45) años

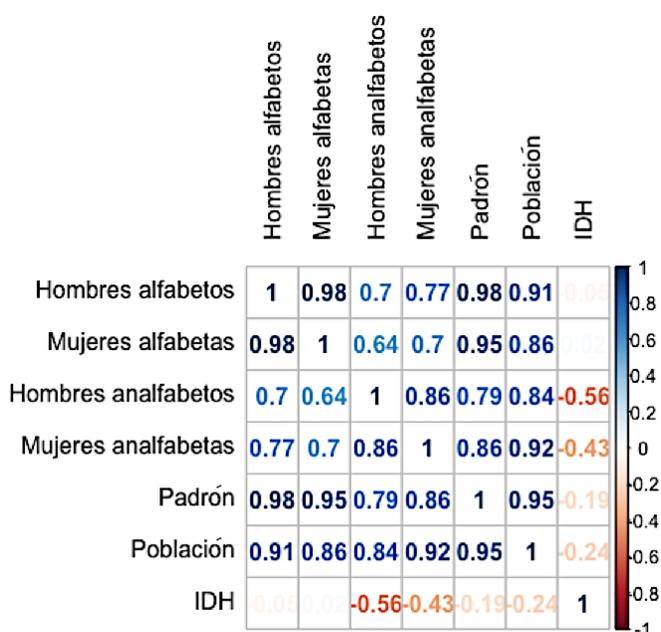
E4: personas inscritas en el padrón electoral entre (46-55) años

E5: personas inscritas en el padrón electoral mayores a 55 años

P: es la proporción de abstinentes (variable respuesta)

Posteriormente se analizó el nivel de asociación de las variables del padrón electoral, el índice de desarrollo humano y el total de la población, a nivel domiciliar ya que es el dato más pequeño territorialmente en el que es expresado el índice anteriormente mencionado.

Figura 11. **Matriz de correlación entre las variables del padrón vs variables de la población**



Fuente: elaboración propia.

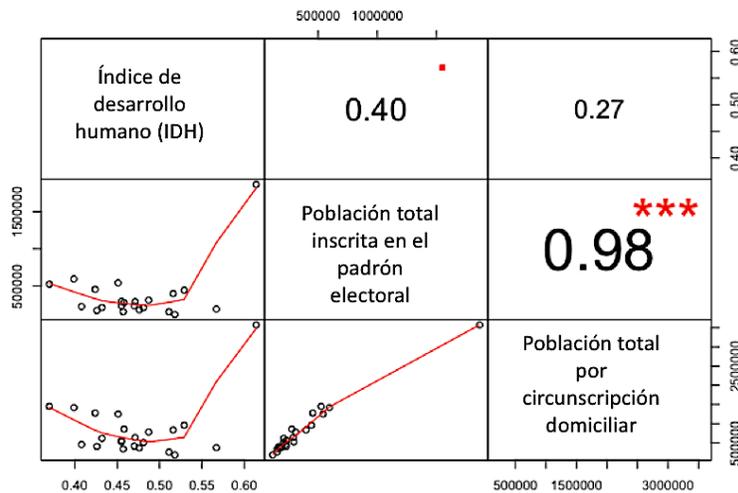
Donde:

IDH es el índice de desarrollo humano.

De conformidad a lo expresado en la figura 11 las variables del padrón electoral y de la población tienen una correlación positiva fuerte, mientras que la asociación de estas con el IDH es negativa y débil con respecto de las demás.

A continuación, se muestra la figura donde se asocian las variables a nivel domiciliar de índice de desarrollo humano, el total del padrón electoral y población total dentro de la misma circunscripción.

Figura 12. **Correlación de Spearman, IDH vs padrón electoral vs población domiciliar**



Fuente: elaboración propia.

Como denota la figura anterior la correlación del índice de desarrollo humano es positiva y relativamente baja con respecto de la población electoral y la población total, siendo estas de cuarenta por ciento y veintisiete por ciento respectivamente, sin embargo la población electoral y la población total tienen un alto grado de asociación positiva, siendo esta de noventa y ocho por ciento.

3.4. Análisis de la influencia de las variables que componen el padrón electoral

Para determinar si la circunscripción domiciliar de cada una de las variables tiene distinción entre sí, se elaboró una tabla con catorce mil novecientas observaciones correspondientes a los tres cientos cuarenta municipios, con observaciones en ambos sexos, alfabetos y analfabetos, y once rangos etarios.

Luego de inferir los empadronados a través de sus respectivas circunscripciones domiciliarias, a través de un modelo se procedió a evaluar los supuestos del mismo, a efecto de determinar la normalidad y homocedasticidad de los datos.

Tabla XV. Prueba de normalidad de Lilliefors (Kolmogorov-Smirnov)

Data	Residuos del modelo
Estadígrafo de prueba (D)	0.30282
Valor de p	$< 2.20000 \times 10^{-16}$

Fuente: elaboración propia.

En este caso el padrón electoral no tiene un comportamiento normal.

Tabla XVI. Prueba de Breusch-Pagan estudentizada

Data	Residuos del modelo
Estadígrafo de Breusch-Pagan (BP)	518.97
Valor de p	$< 2.20 \times 10^{-16}$

Fuente: elaboración propia.

Tabla XVII. **Evaluación de los supuestos del modelo lineal utilizando la prueba global de 4 grados de libertad**

	Valor de p	Decisión
Estadística global	1.131 x 10 ⁸	supuesto no cumple
Oblicuidad	7.499 x 10 ⁵	supuesto no cumple
Curtosis	1.124 x 10 ⁸	supuesto no cumple
Función de enlace	-2.327 x 10 ⁻¹⁰	supuesto cumple
Heteroscedasticidad	1.374 x 10 ³	supuesto no cumple

Fuente: elaboración propia.

Dado que los datos no presentan ni normalidad ni homocedasticidad se procedió a realizar un test no paramétrico de Friedman para comparar las circunscripciones domiciliarias.

Tabla XVIII. **Suma de rangos del Test de Friedman**

Departamentos	Empadronados	r
Alta Verapaz	1892	748
Baja Verapaz	2772	352
Chimaltenango	3652	704
Chiquimula	4532	484
El Progreso	5412	352
Escuintla	6292	616
Guatemala	7172	748
Huehuetenango	8052	1452
Izabal	8932	220
Jalapa	9812	308
Jutiapa	10692	748
Petén	11572	616
Quetzaltenango	12452	1056
Quiché	13332	924
Retalhuleu	14212	396

Continuación de tabla XVIII.

Sacatepéquez	15092	704
San Marcos	15972	1320
Santa Rosa	16852	616
Sololá	17732	836
Suchitepéquez	18612	924
Totonicapán	19492	352
Zacapa	20372	484

Fuente: elaboración propia.

Tabla XIX. **Ajuste de empates en el Test de Friedman**

Valor crítico	16800.000
Valor de p Chi-cuadrado	0.000
Valor F	4604.762
Valor de p en F	0.000

Fuente: elaboración propia

En la tabla XVIII se presentan los grupos homogéneos generados por medio de la prueba post hoc, cada uno denotado por una letra distinta. Si los departamentos tienen la misma letra son significativamente diferentes.

Tabla XX. **Grupos según la probabilidad de diferencias de tratamiento y nivel alfa (0.05)**

Departamentos	Suma de rangos	Grupos
Zacapa	20372	a
Totonicapán	19492	b
Suchitepéquez	18612	c
Sololá	17732	d
Santa Rosa	16852	e
San Marcos	15972	f

Continuación de tabla XX.

Retalhuleu	14212	h
Quiché	13332	i
Quetzaltenango	12452	j
Petén	11572	k
Jutiapa	10692	l
Jalapa	9812	m
Izabal	8932	n
Huehuetenango	8052	o
Guatemala	7172	p
Escuintla	6292	q
El Progreso	5412	r
Chiquimula	4532	s
Chimaltenango	3652	t
Baja Verapaz	2772	u
Alta Verapaz	1892	v

Fuente: elaboración propia.

Ninguna de las circunscripciones domiciliarias tiene similitud entre sí.

3.5. Análisis del mejor ajuste

Se tomaron en cuenta para el ajuste del mejor modelo la evaluación de los estimadores de ajuste: el error cuadrado medio RSME por sus siglas en inglés, el coeficiente de determinación R^2 , y la media absoluta del error MAE por sus siglas en inglés, de los siguientes modelos: *linear Regression*, *kNN*, *Random forest*, *CART*, *Generalized Additive Model using LOESS*, *Linear Regression with Forward Selection*, *Linear Regression with Backwards Selection*, *Linear Regression with Stepwise Selection*, *Multivariate Adaptive Regression Spline*, *Bayesian Ridge Regression*, *Robust Linear Model*, *Projection Pursuit Regression*, *Principal Component Analysis*, *Independent Component*

Regression, eXtreme Gradient Boosting, Support Vector Machines, a pesar de que en el análisis de la normalidad se determina que el comportamiento es no paramétrico se comparan algunos modelos paramétricos para dejar clara constancia de la poca capacidad de ajuste en virtud de la normalidad expresada anteriormente.

A continuación se presenta una serie de tablas donde se corren los dieciséis distintos modelos en la data, con nueve segmentaciones de la misma y todas las variables en su conjunto, es decir, una con todas las variables de la data, dos tablas tomando en cuenta únicamente a las personas inscritas en el padrón electoral de sexo masculino y femenino, dos tablas tomando en cuenta a las personas inscritas en el padrón electoral alfabetas y analfabetas y cinco tablas tomando en cuenta los rangos etarios de cada de las personas inscritas en el padrón electoral, para hacer un total de diez tablas.

Tabla XXI. Modelos evaluados para todas las personas

Modelo	RMSE	R ²	MAE
eXtreme Gradient Boosting	0.06038446	0.61997040	0.04254099
Random forest	0.06352630	0.57553410	0.04452399
Multivariate Adaptive Regression Spline	0.06789412	0.54644660	0.04711718
kNN	0.07238275	0.46118310	0.05026805
CART	0.07264616	0.47020870	0.05352597
Projection Pursuit Regression	0.07370147	0.43747880	0.05025603
Bayesian Ridge Regression	0.08660407	0.25168310	0.06037304
Linear Regression with Stepwise Selection	0.08876864	0.23460860	0.06297067
Principal Component Analysis	0.08895042	0.20360480	0.06460500
Linear Regression with Forward Selection	0.08950268	0.24254000	0.06334864
Linear Regression	0.09142829	0.24545320	0.06194559
Linear Regression with Backwards Selection	0.09412490	0.17719200	0.06596203
Robust Linear Model	0.09423558	0.24210924	0.06166044
Support Vector Machines	0.09772272	0.24048680	0.06141770
Independent Component Regression	0.10030390	0.15659450	0.06810939
Generalized Additive Model using LOESS	4910664217	0.15860470	3292122269

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo de eXtreme Gradient Boosting, el cual cuenta con el número más bajo del estimador raíz del error cuadrado medio, el coeficiente de determinación más alto, y la media del error absoluto. Por lo que el mejor modelo estimado es eXtreme Gradient Boosting, sugerido por los tres estimadores evaluados.

Tabla XXII. Modelos evaluados para personas alfabetas

Modelo	RMSE	R ²	MAE
eXtreme Gradient Boosting	0.08207410	0.29424850	0.06216412
Random forest	0.08337421	0.29133290	0.06207961
Projection Pursuit Regression	0.08533571	0.25098140	0.06232674
kNN	0.08663845	0.22799040	0.06553484
CART	0.08815153	0.22287440	0.06530134
Principal Component Analysis	0.09059404	0.13495240	0.06684077
Bayesian Ridge Regression	0.09145145	0.14908970	0.06683887
Independent Component Regression	0.09151356	0.13294031	0.06681020
Support Vector Machines	0.09182433	0.14654560	0.06652868
Linear Regression with Stepwise Selection	0.09212883	0.14432640	0.06689084
Linear Regression with Forward Selection	0.09327737	0.12237700	0.06768476
Linear Regression with Backwards Selection	0.09356094	0.15684950	0.06712319
Multivariate Adaptive Regression Spline	0.10294590	0.25715200	0.06679739
Robust Linear Model	0.10820770	0.12976053	0.07006565
Linear Regression	0.11075340	0.10759020	0.07052087
Generalized Additive Model using LOESS	675363.80	0.10091860	459197.40

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo de eXtreme Gradient Boosting el cual cuenta con el número más bajo del estimador raíz del error cuadrado medio, el coeficiente de determinación más alto, correspondiente al mismo modelo, y la media del error correspondiente al modelo Random forest. Por lo que el mejor modelo estimado es eXtreme Gradient Boosting, sugerido por dos de tres estimadores evaluados.

Tabla XXIII. Modelos evaluados para personas analfabetas

Modelo	RMSE	R ²	MAE
eXtreme Gradient Boosting	0.07061619	0.47703800	0.05082325
Projection Pursuit Regression	0.07957878	0.35863740	0.05400945
Multivariate Adaptive Regression Spline	0.08059234	0.34006030	0.05463282
Generalized Additive Model using LOESS	0.08148863	0.36847560	0.05485035
kNN	0.08168604	0.32252630	0.05519089
Random forest	0.08235658	0.29785760	0.06195246
CART	0.08337737	0.29416430	0.05696089
Linear Regression with Backwards Selection	0.09081872	0.16357050	0.06606966
Robust Linear Model	0.09127676	0.19563250	0.06416590
Bayesian Ridge Regression	0.09184906	0.16366600	0.06616312
Independent Component Regression	0.09201053	0.14478090	0.06715853
Support Vector Machines	0.09204956	0.18351530	0.06441318
Linear Regression with Forward Selection	0.09225281	0.16659950	0.06611001
Principal Component Analysis	0.09238649	0.16859780	0.06702623
Linear Regression	0.09267382	0.19949320	0.06518250
Linear Regression with Stepwise Selection	0.09306223	0.17082960	0.06661916

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo de eXtreme Gradient Boosting el cual cuenta con el número más bajo del estimador raíz del error cuadrado, el coeficiente de determinación más alto, y la media del error absoluto. Por lo que el mejor modelo estimado es eXtreme Gradient Boosting, sugerido por los tres estimadores evaluados.

Tabla XXIV. Modelos evaluados, para personas de sexo masculino

Modelo	RMSE	R ²	MAE
eXtreme Gradient Boosting	0.06391534	0.57032990	0.04620024
Random forest	0.06429853	0.57566160	0.04663788
Multivariate Adaptive Regression Spline	0.07077819	0.49733840	0.04911984
kNN	0.07198506	0.46565650	0.05206783
Projection Pursuit Regression	0.07381473	0.44426990	0.05064240
CART	0.07634448	0.41358320	0.05365955
Generalized Additive Model using LOESS	0.08522161	0.41068220	0.05196800
Bayesian Ridge Regression	0.08789606	0.26374620	0.06197936
Principal Component Analysis	0.08856000	0.21982676	0.06342098
Linear Regression with Forward Selection	0.08925114	0.25423090	0.06188181
Linear Regression with Backwards Selection	0.08940200	0.22512980	0.06259216
Linear Regression with Stepwise Selection	0.08952810	0.23267040	0.06159175
Support Vector Machines	0.09059800	0.28082940	0.06156493
Linear Regression	0.09064044	0.21790810	0.06236310
Robust Linear Model	0.09160057	0.23165025	0.06183196
Independent Component Regression	0.09614663	0.15963800	0.06803423

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo de eXtreme Gradient Boosting el cual cuenta con el número más bajo del estimador raíz del error cuadrado medio, el coeficiente de determinación más alto, correspondiente al modelo Random forest y la media del error absoluto correspondiente al modelo potenciación del gradiente. Por lo que el mejor modelo estimado es eXtreme Gradient Boosting, sugerido por dos de tres estimadores evaluados.

Tabla XXV. Modelos evaluados, para personas de sexo femenino

Modelo	RMSE	R ²	MAE
Multivariate Adaptive Regression Spline	0.06781591	0.52541330	0.04997210
Random forest	0.06845220	0.51909790	0.04986851
eXtreme Gradient Boosting	0.06983965	0.48885010	0.05030349
kNN	0.07483406	0.41449510	0.05363541
Projection Pursuit Regression	0.07785313	0.37351440	0.05453358
CART	0.08045745	0.36453520	0.05732751
Principal Component Analysis	0.08753232	0.22950610	0.06256617
Linear Regression with Forward Selection	0.08791609	0.24240440	0.06237236
Linear Regression	0.08851553	0.21609950	0.06294220
Linear Regression with Stepwise Selection	0.08939700	0.21873243	0.06313581
Robust Linear Model	0.09055341	0.20550491	0.06288538
Support Vector Machines	0.09086159	0.21826520	0.06227501
Bayesian Ridge Regression	0.09179450	0.21200490	0.06373392
Linear Regression with Backwards Selection	0.09257840	0.21867140	0.06353266
Independent Component Regression	0.09715644	0.15195660	0.06765993
Generalized Additive Model using LOESS	1.90095000	0.34332520	0.78647340

Fuente: elaboración propia

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo Multivariate Adaptive Regression Spline, el cual cuenta con el número más bajo del estimador raíz del error cuadrado medio, el coeficiente de determinación más alto, correspondiente al mismo modelo y la media del error absoluto corresponde al modelo de Random forest. Por lo que el mejor modelo estimado es Multivariate Adaptive Regression Spline, sugerido por dos de tres estimadores evaluados.

Tabla XXVI. Modelos evaluados, para personas entre 18 y 25 años

Modelo	RMSE	R ²	MAE
eXtreme Gradient Boosting	0.07274842	0.45750680	0.05270449
Multivariate Adaptive Regression Spline	0.07344864	0.44802320	0.05335665
Random forest	0.07557343	0.40871120	0.05489868
kNN	0.07755060	0.37705260	0.05549595
Projection Pursuit Regression	0.07794379	0.37231650	0.05514626
CART	0.07941666	0.35785170	0.05792039
Principal Component Analysis	0.08753232	0.22950610	0.06256617
Lineal robusto	0.09391801	0.13817636	0.06818365
Bayesian Ridge Regression	0.09445179	0.12880710	0.06835551
Linear Regression with Backwards Selection	0.09579166	0.13813730	0.06809033
Support Vector Machines	0.09625267	0.14059430	0.06843931
Linear Regression with Stepwise Selection	0.09661168	0.13277120	0.06889846
Linear Regression	0.09663865	0.12484800	0.06825895
Linear Regression with Forward Selection	0.09756616	0.14752800	0.06879324
Independent Component Regression	0.09935757	0.15272900	0.06935361
Generalized Additive Model using LOESS	0.10750530	0.32855280	0.06118549

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo de eXtreme Gradient Boosting el cual cuenta con el número más bajo del estimador raíz del error cuadrado, el coeficiente de determinación más alto, y la media del error absoluto. Por lo que el mejor modelo estimado es eXtreme Gradient Boosting sugerido por los tres estimadores evaluados.

Tabla XXVII. Modelos evaluados, para personas entre 26 y 35 años

Modelo	RMSE	R ²	MAE
eXtreme Gradient Boosting	0.06692693	0.52735790	0.04924626
Random forest	0.06756249	0.52436790	0.04861493
Multivariate Adaptive Regression Spline	0.06795515	0.52046660	0.04907398
kNN	0.07339286	0.44655850	0.05090010
CART	0.07349585	0.44508110	0.05351915
Projection Pursuit Regression	0.07556644	0.41342060	0.05257557
Principal Component Analysis	0.08753232	0.22950610	0.06256617
Independent Component Regression	0.09441575	0.15942790	0.06764305
Linear Regression with Stepwise Selection	0.09457778	0.16646780	0.06731436
Linear Regression with Forward Selection	0.09520137	0.14320640	0.06742942
Robust Linear Model	0.09522427	0.14138019	0.06772259
Bayesian Ridge Regression	0.09548799	0.13971060	0.06797246
Linear Regression with Backwards Selection	0.09599226	0.13877610	0.06751642
Linear Regression	0.09605707	0.14680140	0.06755731
Support Vector Machines	0.09636155	0.15588560	0.06746933
Generalized Additive Model using LOESS	0.11600570	0.34967880	0.06174771

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo de eXtreme Gradient Boosting, el cual cuenta con el número más bajo del estimador raíz del error cuadrado medio, el coeficiente de determinación más alto, correspondiente al mismo modelo y la media del error absoluto corresponde al modelo de Random forest. Por lo que el mejor modelo estimado es eXtreme Gradient Boosting, sugerido por dos de tres estimadores evaluados.

Tabla XXVIII. Modelos evaluados, para las personas entre 36 y 45 años

Modelo	RMSE	R ²	MAE
eXtreme Gradient Boosting	0.0657784	0.5497949	0.0482275
Random forest	0.0661113	0.5390743	0.0477487
Multivariate Adaptive Regression Spline	0.0669837	0.5282433	0.0496741
kNN	0.0706980	0.4950508	0.0519575
CART	0.0716926	0.4677149	0.0522833
Projection Pursuit Regression	0.0756547	0.4050536	0.0515935
Principal Component Analysis	0.0875323	0.2295061	0.0625662
Linear Regression	0.0949478	0.1540524	0.0666149
Robust Linear Model	0.0950788	0.1610353	0.0670781
Linear Regression with Stepwise Selection	0.0955406	0.1632362	0.0672326
Bayesian Ridge Regression	0.0956492	0.1473238	0.0677657
Linear Regression with Forward Selection	0.0958320	0.1662696	0.0674024
Independent Component Regression	0.0959604	0.1764708	0.0677423
Support Vector Machines	0.0996085	0.1452849	0.0675436
Linear Regression with Backwards Selection	0.0996863	0.1600940	0.0681672
Generalized Additive Model using LOESS	0.1537872	0.3694064	0.0641676

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo de eXtreme Gradient Boosting el cual cuenta con el número más bajo del estimador raíz del error cuadrado medio, el coeficiente de determinación más alto, correspondiente al mismo modelo y la media del error absoluto correspondiente al modelo de Random forest. Por lo que el mejor modelo estimado es eXtreme Gradient Boosting, sugerido por dos de tres estimadores evaluados.

Tabla XXIX. Modelos evaluados, para personas entre 46 y 55 años

Modelo	RMSE	R ²	MAE
Multivariate Adaptive Regression Spline	0.06950922	0.50890470	0.04918012
eXtreme Gradient Boosting	0.07101762	0.48449760	0.04967449
kNN	0.07295391	0.45543000	0.05121242
Projection Pursuit Regression	0.07476314	0.42162040	0.05117604
Random forest	0.07530170	0.43085700	0.05181804
CART	0.07535992	0.42057550	0.05459813
Principal Component Analysis	0.08753232	0.22950610	0.06256617
Robust Linear Model	0.09583858	0.19456740	0.06725565
Linear Regression	0.09680922	0.16853800	0.06724038
Independent Component Regression	0.09688803	0.19157000	0.06727585
Linear Regression with Backwards Selection	0.09716488	0.16001650	0.06681897
Generalized Additive Model using LOESS	0.09730288	0.17471590	0.06741828
Bayesian Ridge Regression	0.09730288	0.17471590	0.06741828
Linear Regression with Stepwise Selection	0.09749331	0.13858220	0.06762198
Linear Regression with Forward Selection	0.09929472	0.18009860	0.06728551
Support Vector Machines	0.09953766	0.18490950	0.06642791

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo Multivariate Adaptive Regression Spline el cual cuenta con el número más bajo del estimador raíz del error cuadrado, el coeficiente de determinación más alto, y la media del error absoluto. Por lo que el mejor modelo estimado es Multivariate Adaptive Regression Spline, sugerido por los tres estimadores evaluados.

Tabla XXX. Modelos evaluados, para personas mayores de 55 años

Modelo	RMSE	R ²	MAE
Multivariate Adaptive Regression Spline	0.06681204	0.53905030	0.04762389
eXtreme Gradient Boosting	0.06951292	0.49233960	0.04948899
Random forest	0.07108936	0.47546680	0.04953771
Projection Pursuit Regression	0.07570800	0.42202350	0.05282752
kNN	0.07753034	0.38376110	0.05410250
CART	0.07980150	0.37432900	0.05534066
Principal Component Analysis	0.08753232	0.22950610	0.06256617
Linear Regression	0.09539678	0.18918940	0.06642323
Linear Regression with Backwards Selection	0.09559740	0.19532400	0.06653822
Linear Regression with Stepwise Selection	0.09650190	0.16926620	0.06635361
Linear Regression with Forward Selection	0.09722943	0.18278490	0.06669527
Independent Component Regression	0.09838511	0.19215530	0.06705542
Bayesian Ridge Regression	0.09860386	0.16388200	0.06734791
Robust Linear Model	0.11346720	0.17812589	0.06886378
Support Vector Machines	0.12137210	0.20197200	0.06997975
Generalized Additive Model using LOESS	0.34345250	0.29315940	0.09330546

Fuente: elaboración propia.

Como se muestra en la tabla anterior el mejor ajuste corresponde al modelo Multivariate Adaptive Regression Spline el cual cuenta con el número más bajo del estimador raíz del error cuadrado, el coeficiente de determinación más alto, y la media del error absoluto. Por lo que el mejor modelo estimado es Multivariate Adaptive Regression Spline, sugerido por los tres estimadores evaluados.

Tabla XXXI. **Resumen de los mejores modelos de cada data analizada**

Data analizada	Modelo	RSME	R²	MAE
Modelo completo	eXtreme Gradient Boosting	0.06038446	0.61997040	0.04254099
Hombres	eXtreme Gradient Boosting	0.06391534	0.57032990	0.04620024
De 36 a 45 años	eXtreme Gradient Boosting	0.06577841	0.54979490	0.04822745
Mayores a 55 años	Multivariate Adaptive Regression Spline	0.06681204	0.53905030	0.04762389
De 26 a 35 años	eXtreme Gradient Boosting	0.06692693	0.52735790	0.04924626
Mujeres	Multivariate Adaptive Regression Spline	0.06781591	0.52541330	0.04997210
De 46 a 55 años	Multivariate Adaptive Regression Spline	0.06950922	0.50890470	0.04918012
Analfabetas	eXtreme Gradient Boosting	0.07061619	0.47703800	0.05082325
De 18 a 25 años	eXtreme Gradient Boosting	0.07274842	0.45750680	0.05270449
Alfabetas	eXtreme Gradient Boosting	0.08207410	0.29424850	0.06216412

Fuente: elaboración propia.

Como lo muestra la tabla anterior el mejor modelo es eXtreme Gradient Boosting cuando se utilizan todas las variables de los votos emitidos, en virtud de ostentar la mejor estima.

De los tres modelos con mejor ajuste de cada sección, el modelo de Random forest, combina el modelo de arboles predictores en la cual cada árbol donde de un valor representado por un vector aleatorio, probado independientemente con la misma distribución, haciendo una agrupación de arboles no correlacionados para luego promediarlos, a comparación del Multivariate Adaptive Regression Spline que adapta una curva diferenciable definida en porciones mediante polinomios, los cuales utiliza para interpolar, para esto es necesario polinomios de bajo grado, evitando las oscilaciones, obteniendo resultados similares a polinomios de grado alto.

El modelo de eXtreme Gradient Boosting utiliza algoritmos que optimiza funciones, esta se elije a través de su elección iterativa que señala en dirección de un gradiente negativo. Está diferencia entre los modelos evaluados hace que este último sea el que mayor alcance ostenta para este caso en específico.

4. DISCUSIÓN DE RESULTADOS

4.1. Análisis descriptivo previo

La data que contiene el conteo de los votos por zonas municipales, la cual cuenta con quinientas ochenta observaciones, tiene un rango alto entre cada una de las variables observadas, es decir, que hay observaciones con muy pocos datos y observaciones con un alto conteo de votos por lo que existe una alta variación entre los mismos, razón descrita en la columna de la desviación estándar. Debido a la distancia entre los datos el error estándar aumenta a medida de la variación que ostentan las variables, por lo que el grupo de personas analfabetas al tener menos distancia entre los datos tienen un menor error estándar.

Esta variación entre las observaciones, hace que existan valores extremos por lo que para una descripción de los datos se calcula la media ajustada, que excluye a estos valores, la desviación absoluta de la media expresa la distancia promedio entre cada valor y la media calculada, la cual es relativamente grande concatenando una alta variación entre las observaciones. Cabe destacar que los datos están sesgados a la izquierda y todas las variables son leptocúrticas, en mayor magnitud las personas alfabetas.

La proporción de votos emitidos, que cuenta igualmente con quinientas ochenta observaciones, tiene un rango aparentemente pequeño entre cada una de las ellas, dada la naturaleza porcentual de la misma, es decir que hay observaciones con muy poca o muy alta abstención, no pueden ser

dimensionadas a efecto que estas dependen de una ponderación derivada de la magnitud y participación de una observación en el conteo total de votos, por lo que existe una poca variación entre los mismos, razón descrita en la columna de la desviación estándar, sin embargo esta no detalla que la variable respuesta este compactada. Debido a la distancia entre los datos es pequeña el error estándar se aproxima a cero.

Esta variación en la proporción de votos emitidos, hace que no existan valores extremos por lo que para una descripción de los datos se calcula la media ajustada, que excluye a estos valores, dando un valor similar, la desviación absoluta de la media expresa la distancia promedio entre cada valor y la media calculada, lo cual define mejor aunque no de manera precisa una variación existente en el fenómeno evaluado. Cabe destacar que la variable respuesta, en este caso la proporción de votos emitidos está sesgada a la derecha y es mesocurtica.

Como aproximación a la normalidad de los datos se ajustó la variable respuesta, a una distribución de Weibull ya que siendo ésta analizada con el software estadístico se determinó a través de los criterios de información bayesiana y de Akaike que la distribución antes mencionada es la mejor estimada a efecto de determinar la forma en la que se distribuye dicha variable.

La data que contiene el padrón electoral, la cual cuenta con cuatrocientas cuarenta observaciones, ostenta alto rango entre cada una de las variables estudiadas, es decir, que hay observaciones con muy pocos datos y otras con un alto número de empadronados por lo que existe una alta variación entre los mismos, razón descrita en la columna de la desviación estándar. Debido a la distancia entre los datos, el error estándar aumenta a medida de la variación que

ostentan las variables, por lo que el grupo de analfabetas, al tener menos distancia entre los datos, tienen un menor error estándar.

En relación a lo anteriormente expuesto, dichas variaciones hacen que existan valores extremos, por lo que para una descripción de los datos se calcula la media ajustada, que excluye tales valores, la desviación absoluta de la media expresa la distancia promedio entre cada valor y la media calculada, lo cual es relativamente grande concatenando una alta variación entre las observaciones. Cabe destacar que los datos están sesgados a la izquierda y todas leptocurticas, en mayor magnitud las personas alfabetas.

4.2. Análisis de normalidad

De conformidad con lo que establece el test de normalidad de Shapiro-Wilk se denota que este valor es cercano a cero por lo que se rechaza la hipótesis nula afirmando que la variable respuesta la proporción de votos emitidos, no obedece a un comportamiento normal.

Luego de esto el modelo de regresión lineal generó unos residuos de tal forma que su comportamiento no obedece a la normalidad como se denota en la Tabla X en la Figura 7, donde el estadístico de prueba de normalidad de Shapiro-Wilk sobre los residuos del mismo, tiene un valor inferior al cinco por ciento de igual manera en la Tabla IX y la Figura 7 donde se realiza la prueba estudentizada de Breusch-Pagan se deduce que el comportamiento de los residuos son heterocedásticos, por lo que no cumple con los supuestos mínimos para ser descrito con una recta. Como se muestra en la Tabla XI de la evaluación de los supuestos del modelo lineal usando la prueba global, la cual no cumple ninguno

de los supuestos de estadístico global, agudez, curtosis, función de enlace heterocedasticidad.

Sin embargo, estos supuestos tienen un carácter preliminar debido a que se deben agotar todas aquellas instancias, que para este caso son las transformaciones matemáticas, que se aplican sobre la variable respuesta a efecto de probar que el conjunto de datos bajo análisis se puede inferir con una recta.

Por consiguiente como se muestra en la tabla XII se corrieron las diferentes transformaciones las cuales fueron estimadas a través de la agudez, la curtosis, prueba de normalidad de Shapiro-Wilk y test de Breusch-Pagan para la evaluación de la homocedasticidad, de las cuales ninguna tuvo como resultado una evaluación estadística significativa, que pudiera inferir los datos de manera lineal.

Cabe destacar que se suprimieron de la evaluación las transformaciones de desplazamiento y de raíces cuadradas, ya que el rango de λ predeterminado para las transformaciones, es muy corto por la naturaleza numérica de la misma, a pesar de que hubieron transformaciones que sí se pudieron realizar ninguna de estas tuvo como resultado una estimación normal satisfactoria, como se muestra en la tabla XII.

En virtud de lo anteriormente expuesto se puede afirmar que se agotaron los procesos en la búsqueda de la normalidad de la data bajo estudio, por lo que ninguna de las once transformaciones posibles, tiene un nivel de significancia suficiente para ser considerado normal, determinando que la data bajo estudio debe ser analizada de forma no paramétrica.

4.3. Análisis de correlación

Como se menciona en el subtítulo anterior el comportamiento de los datos es no paramétrico por lo cual el estadístico de prueba para la correlación fue el estadígrafo de Spearman.

Tal como se muestra en la figura IX las variables de la data que expresan una cantidad de votos de las personas inscritas en el padrón electoral con la característica de alfabetismo, tienen una estrecha asociación entre sí, independientemente de los rangos etarios y el sexo de las mismas. Cabe destacar entre este grupo de personas alfabetas y el grupo de edad mayor, tanto como para los de género masculino como para los de género femenino, tienen una correlación mucho menor que con sus contrapartes más jóvenes, ya que el coeficiente de correlación de Spearman oscila entre ochenta y cien por ciento, teniendo los valores más altos en los rangos de menor edad.

Es oportuno mencionar que el alfabetismo en todos sus niveles se correlacionan positivamente entre sí, sin embargo se correlaciona negativamente con la proporción de votos emitidos, de una manera tenue teniendo una correlación desde menos punto veinte por ciento hasta menos cuarenta y cuatro por ciento.

El analfabetismo tiene una correlación entre setenta y cien por ciento cuando se asocia entre sí, en los diferentes rangos etarios y el sexo. Este mismo grupo tiene una correlación positiva entre el cuarenta y sesenta por ciento, con la variable respuesta, cabe destacar que es el grupo del padrón electoral con nivel de correlación más alto.

Este nivel de correlación de las personas analfabetas, se debe a la poca cantidad de personas que integran este segmento, al tener una población relativamente pequeña a su contraparte alfabetada, es más sensible a las fluctuaciones que genera el abstencionismo en el número total de votos emitidos, por lo que el efecto de dicho fenómeno se denota de mejor manera.

En el segmento de personas analfabetas los niveles etarios tienen una relación positiva, de la misma manera que con las personas de sexo masculino alfabetadas de menor edad, y más tenue entre las las personas alfabetadas de sexo masculino de mayor edad, el cual oscila entre el setenta y cien por ciento.

Este mismo fenómeno se replica cuando las personas analfabetas se asocian con las mujeres alfabetadas teniendo una muy poca o casi nula asociación con las mujeres jóvenes y teniendo una asociación negativa mucho mayor a medida que se compara a lo largo del rango etario de las mujeres alfabetadas. El nivel de asociación que tienen las personas analfabetas inscritas en el padrón electoral con la variable respuesta es positiva y va desde el cuarenta y nueve por ciento hasta el sesenta y tres por ciento.

De acuerdo con la matriz de correlación, existe una diferencia en función de las personas, independientemente del sexo tienen un nivel de asociación similar al igual que los niveles etarios es únicamente el alfabetismo el que tiene una distinción aparente, esto puede dar lugar a conjeturas desviadas de la realidad porque se debe tomar en cuenta que nuestra variable respuesta, la proporción de votos emitidos y las variables que utilizamos a efecto de describir el comportamiento de la misma, se realiza a través de un conteo de votos, por lo que la situación actual de la República de Guatemala en los eventos electorales genera amplia distancia de la cantidad de votos emitidos, entre

personas analfabetas y alfabetas, siendo estas últimas las que representan un porcentaje superior al noventa por ciento de la participación total.

4.4. Análisis de la influencia de las variables que componen el padrón electoral

El test de Friedman con el que se analizó la base de datos del padrón electoral, muestra una clara distinción entre los grupos que la componen, tomando en cuenta que la prueba realizada se ejecutó a un nivel de confianza del noventa y cinco por ciento, por lo que se debe obtener un p valor superior al cinco por ciento para interpretar los efectos de cada grupo como no significativos, derivado que la prueba antes mencionada establece un área bajo la curva denotada por su p valor muy cercana a cero lo que indica que los grupos bajo análisis ostentan diferencias significativas entre sí.

Para tal efecto, se realizaron veintidos grupos uno por cada circunscripción domiciliar, debido a que el grupo más pequeño de análisis de la data que contiene los empadronados totales, tiene como segmento cada uno de los trescientos cuarenta municipios, con el objeto de poder establecer cada uno como observación estos debían de ser agrupados en la clasificación inmediata superior, que para este caso es el nivel departamental.

Es importante mencionar que la clasificación departamental, englobó las participaciones del sexo, alfabetismo y rango etario, dado que estas conjuntamente con los municipios obedecen a una misma clasificación, por lo tanto la clasificación departamental, denota el contenido del total de empadronados.

En virtud de lo expresado en las tablas, XIII, XIV y XV, la data bajo análisis no cumple con los criterios para ser determinada parametricamente, razón por la cual el análisis de varianza se realiza a través del test de Friedman.

De conformidad a la tabla XVII, luego de realizar el Test de Friedman, se estableció que los rangos de las circunscripciones dominiciares, tienen diferencias significativas, a efecto de que el p valor del test es cercano a cero, por lo que no se alcanza el cinco por ciento requerido, para determinar lo contrario.

Dado que la prueba de Friedman tiene diferencias significativas se realizó un análisis post hoc para agrupar de acuerdo a la probabilidad de la diferencia de las participaciones, la pretención de estos análisis es determinar si existen grupos con semejanzas suficientes para calificar sus diferencias estadísticamente como no significativas.

Al aplicar la prueba post hoc de Friedman, como se muestra en la tabla XVIII cada una de las circunscripciones domiciliarias en las que se divide el padrón electoral tienen un grupo distinto, es decir que cada uno de los departamentos posee una variación de tal magnitud que la hace distinta de los demás por lo que existen diferencias significativas que distinguen a cada uno de los grupos llegando así a determinar que no existen departamentos iguales de conformidad a la participación de cada uno en la integración del mismo.

Como anteriormente se describe, los rangos entre cada uno de los departamentos, no logran obtener una semejanza estadística ya que cada uno en comparación con cada uno de sus homólogos ostenta una diferencia significativa, en virtud de tal situación se hace imposible aglomerar a más de dos

departamentos en un grupo, quedando cada circunscripción en un grupo individual.

4.5. Análisis del mejor ajuste

En la modelización de los datos se compararon modelos lineales, multivariados y no paramétricos, entre otros, con el objeto de obtener la mejor bondad de ajuste a través de los estimadores, para lo cual se corrieron dieciséis modelos, a efecto de poder comparar los mejores ajustes, estos se modelaron infiriendo la proporción de abstinentes en un proceso electoral, con base en la cantidad de votos emitidos.

Como se observa en las tablas de la XIX a la XXVIII, los modelos ejecutados con una parte de la data tienen una estimación inferior al modelo que utiliza todas las variables independientes, es decir que cada una de las variables independientes aporta información a tal punto que la ausencia de una de ellas en el modelaje disminuye los estimadores del modelo.

Concatenando lo anteriormente descrito, la proporción de votantes se infiere de manera más precisa cuando se analiza con todas las variables, esto denota que la abstención afecta a cada conglomerado por lo que la ausencia de cada uno de estos, hace que el modelo pierda información, explicando en menor medida la varianza del fenómeno.

En el análisis individual que se realizó a cada una de las variables segmentadas, se tienen estimadores distintos derivado que cada una de estas segmentaciones tiene distinta variación, por lo que al tener una forma regular en

el comportamiento de las segmentaciones antes mencionadas tendría un mejor ajuste.

Como se muestra en la tabla XXIX los segmentos de la data que mejor bondad de ajuste tienen al ser analizados de forma individual, es el grupo de las personas de sexo masculino que emiten voto y el segmento con mayor variación que expresa el menor ajuste es de las personas alfabetas que emitieron su voto.

El conglomerado de las personas alfabetas en comparación con el de las analfabetas, tiene una cantidad muy distante de votos emitidos por lo que los votos que emiten las personas alfabetas, están propensas a tener una mayor variación a lo largo de las zonas urbanas y rurales de cada circunscripción municipal dentro de la república, por consiguiente, ese caudal tan amplio aumenta la variación entre las observaciones limitando un ajuste de sumas ponderadas como lo es la regresión. Por otra parte, un segmento que mantiene la variación entre cada una de las observaciones mejora estas sumas ponderadas haciendo que el ajuste sea mucho mayor.

Los modelos que fueron evaluados, incluyeron modelos paramétricos, no paramétricos y multivariados, a efecto de comparar el desempeño de cada uno de ellos, por consiguiente, de acuerdo a Ramírez (2013), los modelos multivariados describen de mejor manera el fenómeno del abstencionismo para este evento en especial.

Dentro de los diez segmentos en los que se analizó la data destacan tres modelos principales en los cuales el fenómeno expresó su mejor ajuste, los cuales son: Random forest, Multivariate Adaptive Regression Spline y eXtreme Gradient Boosting, siendo este último el que mejor bondad de ajuste mostró.

Entre los segmentos referentes a la variable sexo de la data, se establece que las personas de sexo masculino tienen un mejor ajuste que las de sexo femenino, en virtud que los hombres para el comicio evaluado tienen un comportamiento constante en todos los niveles etarios tanto alfabetos como analfabetos, a comparación de las mujeres, derivado de esto la proporción de votos emitidos por hombres tiene una mayor capacidad explicativa que las mujeres.

Con respecto al alfabetismo, el sector de votos emitidos por analfabetos tiene mayor capacidad explicativa que el de alfabetos, 0.47 y 0.29 de la varianza explicada respectivamente, esto se debe a la variación que existe en el sector de alfabetos, ya que este siendo el que tiene una participación electoral superior al noventa por ciento, la variabilidad de los rangos etarios y el sexo en la emisión de votos se refleja en el sector de alfabetos, en virtud de que este voto se concentra en zonas urbanas de mayor participación.

En cuanto a la edad, el segmento con mayor capacidad predictiva es el grupo de las personas entre 36 a 45 años, siendo este el mejor ajustado seguido por el de las personas mayores 55 años, 26 a 35 años, 46 a 55 años respectivamente, estos cuatro rangos etarios tienen poca diferencia entre sí, debido a que sus estimadores tienen magnitudes muy cercanas por lo que se deduce que el comportamiento de estos cuatro rangos etarios no tienen tanta variación entre sus observaciones como la del primer rango etario que comprende a las personas más jóvenes del padrón estando entre 18 a 25 años.

Este primer rango etario es tan disperso a lo largo de todas las observaciones que tiene poca capacidad predictiva con respecto de sus homólogos, siendo así el segmento etario con menor bondad de ajuste.

CONCLUSIONES

1. El modelo de mejor ajuste con base a los estimadores de la raíz del error cuadrado medio, el coeficiente de determinación y la media del error absoluto el de eXtreme Gradient Boosting, ya que este fue el mejor estimado dentro del análisis explicando con el menor error posible el sesenta y dos por ciento de la varianza del fenómeno.
2. La asociación entre el IDH y las variables del padrón electoral como el sexo y el alfabetismo, correlacionadas a nivel domiciliar, no presentan ningún nivel relevante, caso contrario de la vinculación estadística entre el total de la población, el padrón electoral y sus variables civiles, y la ausencia o no de analfabetismo, sin embargo esta asociación se debe a que existe un constante aumento de la población, concatenado las resultas del test de Friedman se considera que las variables referidas presentan diferencias significativas a nivel domiciliar a pesar de la vinculación expresada.
3. De acuerdo con los modelos de regresión, se determina que todas las variables y sus permutas, entre sexo, abstencionismo y rangos etarios son influyentes dentro del modelo derivado de que la ausencia de de una de ellas conlleva a la pérdida de información de tal forma, que la estima del modelo se ve afectada.
4. Luego de la modelización de los datos el modelo mejor evaluado basado en los estimadores, el modelo eXtreme Gradient Boosting, inferido con todas las variables independientes, se establece como el modelo de mejor ajuste.

RECOMENDACIONES

1. Se anima al Tribunal Supremo Electoral para que a través del registro de ciudadanos y el departamento de informática amplíen el registro de los votantes, presisando la data obtenida, derivado que con el tiempo se necesita precisión para poder obtener un análisis más profundo con respecto al comportameinto de las variables.
2. Dar seguimiento y segmentar la data de conformidad a los períodos que la ley establece para un mejor análisis, es decir que los rangos etarios pueden establecerse cada cuatro años en virtud de que las elecciones generales se efectúan en ese mismo período de tiempo.

REFERENCIAS

1. Acuerdo número 018-2007. *Reglamento de la Ley Electoral y de Partidos Políticos*. Tribunal Supremo Electoral. Diario de Centroamérica. Guatemala 23 de enero de 2007. Recuperado de <https://www.tse.org.gt/images/LEPP/reglamento.pdf>
2. Alonso, J. y Montenegro, S. (2015). Estudio de Monte Carlo para comparar 8 pruebas de normalidad sobre residuos mínimos cuadrados ordinarios en presencia de procesos autorregresivos de primer orden. *Estudios Gerenciales*. (31) 253-265 doi: 10.1016/j.estger.2014.12.003
3. Burnham, K. P. y Anderson, D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Estados Unidos: Springer-Verlag.
4. Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Estados Unidos: Wiley.
5. Canavos, G. C. (1988) *Probabilidad y estadística*. México: McGrall-Hill.
6. Chavarría Mora, E. (2014). Relación entre desarrollo humano y abstencionismo a nivel local en Costa Rica para las elecciones del 2010. *Revista de Derecho Electoral*. (18). Recuperado de <https://dialnet.unirioja.es/servlet/articulo?codigo=5607359>

7. Chai, T. y Draxler, T. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literatura. *Geoscientific Model Development*. (7), 1247-1240. doi:10.5194/gmd-7-1247-2014
8. Constitución Política de la República de Guatemala. Asamblea Nacional Constituyente. Diario de Centroamérica. Guatemala 14 de enero de 1986. Recuperado de https://www.oas.org/juridico/mla/sp/gtm/sp_gtm-int-text-const.pdf
9. Decreto 1-85. *Ley Electoral y de Partidos Políticos*. Asamblea Nacional Constituyente. Diario de Centroamérica. Guatemala 25 de noviembre de 2016. Recuperado de <https://www.tse.org.gt/images/LEPP.pdf>
10. Del Valle Moreno, J. y Guerra Bustillo, W. (2012). Universidad Agraria de la Habana. La multicolinealidad en modelos de regresión lineal múltiple. *Revistas de Ciencias Técnicas Agropecuarias*. 4. (21). Recuperado de http://scielo.sld.cu/scielo.php?pid=S207100542012000400013&script=sci_arttext&tlng=pt
11. Fornos, C., Power, T. y Garand, J. (2004). Explaining Voter Turnout in Latin America, 1980 to 2000. *Comparative Political Studies*. 8 (37) 09-40. doi: 10.1177/0010414004267981
12. Hernández, O. (1991). Universidad de Costa Rica. Anuario de Estudios Centroamericanos. Análisis del abstencionismo en las elecciones presidenciales de Costa Rica en el periodo 1953-1986. *Portal de*

Revistas Académicas. 16. (16). Recuperado de <https://revistas.ucr.ac.cr/index.php/anuario/article/view/2734>

13. Instituto Federal Electoral. (1998). Democracia, Educación y Capacitación Cívico – Electoral. *Estudio Sobre Abstencionismo en México.* Recuperado de https://portalanterior.ine.mx/documentos/DECEYEC/vgn_investigacion/estudio_sobre_abstencionismo.htm
14. Kendall, M, G. y Stuart, A. (1967). *The advanced theory of statistics.* Estados Unidos: Hafner Publishing Company
15. Kluwer, W. (2019). *Abstencionismo electoral.* Recuperado de https://guiasjuridicas.wolterskluwer.es/Content/Documento.aspx?params=H4sIAAAAAAAAAEAMtMSbF1jTAAAUNDYwsTtbLUouLM_DxblwMDCwNzAwuQQGZapUt-ckhIQaptWmJOcSoAozpugDUAAAA=WKE
16. Lago Peñas, I. (2000). *Identidades duales y abstención diferencial en las comunidades autónomas: los casos de Cataluña Galicia y País Vasco.* 2 (9) 83-103. Recuperado de https://minerva.usc.es/xmlui/bitstream/handle/10347/7719/pg_085106_dereito9-2.pdf?sequence=1
17. López, J. F. (2017) Coeficiente de determinación (R cuadrado). *Economipedia.* [versión electrónica]. Murcia, España: Economipedia, <https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html>

18. Lutz, B. (2005). Revista mexicana de sociología. *La participación electoral inconclusa: abstencionismo y votación nula en México*. 4 (67). Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-25032005000400005
19. Mohd, N, y Bee, Y. (2011). *Power comparisons of Shapiro-Wilk, Komogorov-Simirnov, Lilliefors and Andreson-Darling tests*. Journal of Statistical Modeling and Analytics. 2(1), 21-23.
20. Nuñez-Colín, C. (2018). Análisis de varianza no paramétrica: un punto de vista a favor para utilizarla. *Acta Agrícola y Pecuaria*. 4 (3), 69-79. doi: 10.30973/aap/2018.4.3/1
21. Pons Sarlot, LI. (2019). *L'abstenció diferencial i el vot dual a les Illes Balears*. Recuperado de <http://hdl.handle.net/10609/91466>
22. Ramírez, R. J. (2013). *Municipales 2012. Indagando la abstención y otros aspectos*. Recuperado de https://archivos.lyd.org/other/files_mf/sip132municipales2012indagandoenlaabstencionyotrosaspectosjramirezfebrero2013.pdf
23. Restrepo, L. y González, J. (2007). De Pearson a Spearman. *Revista Colombiana de Ciencias Pecuarias*. 20 (2), 183-192. Recuperado de: <http://www.scielo.org.co/pdf/rccp/v20n2/v20n2a10.pdf>
24. Riba, C. (2000). Reis: Revista Española de Investigaciones Sociológicas. *Voto dual y abstención diferencial. Un estudio sobre el*

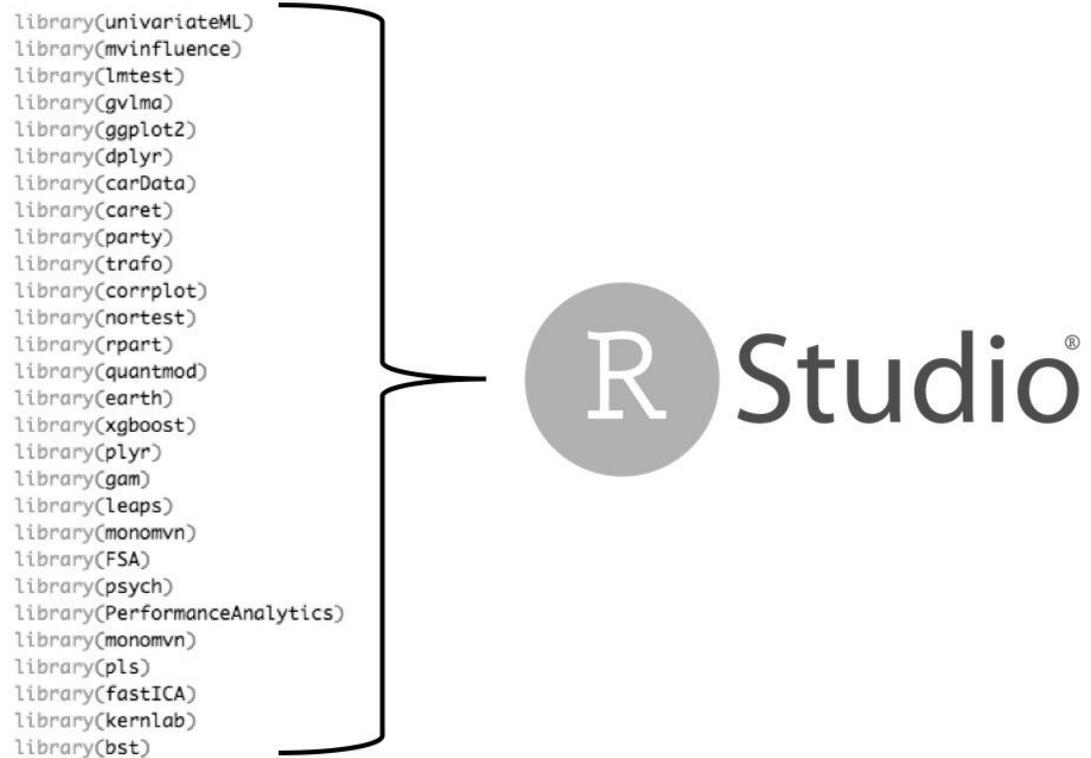
comportamiento electoral en Cataluña. 91, 59-88 doi:
10.2307/40184275

25. Ruiz, M, A., Prado, A. y San Martín, R. (2010). Modelos de ecuaciones estructurales. *Papeles del Psicólogo.* 31 (1), 34-45. Recuperado de: <https://www.redalyc.org/pdf/778/77812441004.pdf>
26. Santana, A., Rama, J., Arellano, R. y Montero, J. R. (2018). *Elecciones parlamentarias Chile 2017: los perfiles de los votantes.* Análisis. Recuperado de <http://host-201-221-123-195.ip.udd.cl/handle/11447/2231>
27. Soto Caballero, H. R. (2013). El abstencionismo decisonal la tensión entre ausencia de participación electoral y ausencia de decisión electoral. (Tesis de maestría). Pontificia Universidad Javeriana, Bogotá, Colombia.
28. Soto Zazueta, I. M. (2016). Dinamismo y diversidad en la ciencia política latinoamericana, VII Congreso de la Asociación Latinoamericana de Ciencia Política. *Abstencionismo y votos nulos en México: un modelo econométrico de sus determinantes.* Recuperado de <https://books.google.es/books?hl=es&lr=&id=KpBcDwAAQBAJ&oi=fnd&pg=PA291&dq=abstencionismo+regresion+lineal&ots=xorJoTu4M2&sig=arK2N1lxI38qahpjCfxQTnY5s#v=onepage&q=abstencionismo%20regresion%20lineal&f=false>
29. Walpole, R., Myers, R., Myers, S., y Ye, K. (2012) *Probabilidad y estadística para ingeniería y ciencias*, México: Pearson Educación de México, S. A.

30. Wooldridge, J. (2010) . *Introducción a la econometría: Un enfoque moderno*. Estados Unidos. Cengage Learning.

APÉNDICES

Apéndice 1. Librerías utilizadas en el software estadístico R Studio para el análisis de datos



Fuente: elaboración propia.

Apéndice 2. **Métodos Utilizados para generar los modelos de regresión de la librería “caret” del programa R Studio**

```
method = "lm"  
method = "knn"  
method = "rf"  
method = "rpart2"  
method = "gamLoess"  
method = "leapForward"  
method = "leapBackward"  
method = "leapSeq"  
method = "earth"  
method = "blassoAveraged"  
method = "rlm"  
method = "ppr"  
method = "pcr"  
method = "icr"  
method = "xgbTree"  
method = "svmLinear"
```



Métodos utilizados para generar los modelos de regresión de la librería “caret” del programa R Studio

Fuente: elaboración propia.

