ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

ANÁLISIS ESTADÍSTICO DE LOS PROCESOS ELECTORALES EN ECUADOR

ANÁLISIS EXPLORATORIO DE ESTRUCTURAS, TENDENCIAS Y RELACIONES ESTADÍSTICAS DE VARIABLES SOCIALES, ECONÓMICAS Y ELECTORALES EN ECUADOR

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN CIENCIAS DE LA COMPUTACIÓN

JONATHAN ALEXANDER HERRERA FLORES

jonathan.herrera01@epn.edu.ec

DIRECTOR: LUIS ENRIQUE MAFLA GALLEGOS

enrique.mafla@epn.edu.ec

DMQ, septiembre 2022

CERTIFICACIONES

Yo, JONATHAN ALEXANDER HERRERA FLORES declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

JONATHAN ALEXANDER HERRERA FLORES

Certifico que el presente trabajo de integración curricular fue desarrollado por JONATHAN ALEXANDER HERRERA FLORES, bajo mi supervisión.

LUIS ENRIQUE MAFLA GALLEGOS

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

JONATHAN ALEXANDER HERRERA FLORES

LUIS ENRIQUE MAFLA GALLEGOS

DEDICATORIA

Este trabajo va dedicado a Dios, apoyándome en él he llegado al final de esta etapa en mi aprendizaje.

“Te haré entender, y te enseñaré el camino en que debes andar;

Sobre ti fijaré mis ojos.”

Salmos 32:8

AGRADECIMIENTO

Quiero agradecer a mi familia por su apoyo y paciencia en cada paso, a mis profesores por haberme instruido en los conceptos de esta maravillosa carrera, a Dios por haberme guiado en todas las decisiones a lo largo de mi carrera, y a mi prometida y futura esposa, por siempre estar ahí cuando más la necesitaba con todo su amor. Alcanzar esta meta únicamente fue posible gracias a todos ellos.

ÍNDICE DE CONTENIDO

RESUMEN

El presente trabajo inició con un análisis exploratorio de datos usando el entorno de Kaggle, mediante el método hipotético-deductivo. Los datos usados provienen del CNE sobre las elecciones generales del 2021 en Ecuador, y de la encuesta de empleo, desempleo y subempleo del INEC. Se realizó observaciones, se planteó dos hipótesis, se dedujo sus consecuencias y luego se intentó refutar las hipótesis. La primera hipótesis planteó que la diferencia observada entre votos registrados y la suma total de votos, es debido a errores comunes cometidos en las juntas. Esta hipótesis no pudo ser refutada con los datos observados, por lo que es tentativamente verdadera según el análisis realizado. La segunda hipótesis planteó que los votantes tendrían cierta preferencia por un candidato, según el estrato socioeconómico al que pertenecen. La misma fue refutada parcialmente ya que, según el análisis, hubo preferencia por un candidato de parte de votantes de estrato bajo, pero no en el caso de los votantes de estrato alto. Adicionalmente, se realizó un análisis computacional general, y una discusión sobre ventajas y desventajas del entorno de Kaggle. Finalmente, se realizó un análisis computacional específico para el comportamiento que tuvo el entorno de Kaggle durante el análisis exploratorio anterior. Este análisis se realizó usando el método hipotético-deductivo, con una hipótesis que preveía las secciones que ocuparían más recursos computacionales. Esta hipótesis no pudo ser refutada por los datos analizados, pero se encontró secciones de código adicionales con alto consumo de recursos computacionales.

PALABRAS CLAVE: análisis exploratorio, análisis computacional, evaluación de rendimiento, procesos electorales, condiciones socioeconómicas.

ABSTRACT

The present work started with an exploratory data analysis using the Kaggle environment, through the hypothetical-deductive method. The data used comes from the CNE on the 2021 general elections in Ecuador, and the INEC employment, unemployment, and underemployment survey. Observations were made, two hypotheses were raised, their consequences were deduced and then an attempt was made to refute the hypotheses. The first hypothesis stated that the difference observed between registered votes and the total sum of votes is due to common errors made in the meetings. This hypothesis could not be refuted with the observed data, so it is tentatively true according to the carried-out analysis. The second hypothesis proposed that voters would have a certain preference for a candidate, depending on the socioeconomic stratum to which they belong. It was partially refuted since, according to the analysis, there was a preference for a candidate on the part of low-stratum voters, but not in the case of high-stratum voters. Additionally, a general computational analysis, and a discussion about the advantages and disadvantages of the Kaggle environment, were performed. Finally, a specific computational analysis was performed for the behavior of the Kaggle environment during the previous exploratory analysis. This analysis was carried out using the hypothetical-deductive method, with a hypothesis that anticipated the sections that would occupy more computational resources. This hypothesis could not be refuted by the analyzed data, but additional code sections with high consumption of computational resources were found.

KEYWORDS: exploratory analysis, computational analysis, performance evaluation, electoral processes, socioeconomic conditions.

1. INTRODUCCIÓN

Este trabajo se enfocó en combinar los conocimientos adquiridos en la carrera de Ingeniería en Ciencias de la Computación, y la aplicación de los mismos. Mediante este componente se combinó varias asignaturas de la malla curricular, para apreciar su importancia individual y colectiva. En este trabajo se realizó tres tipos de análisis, un análisis exploratorio de datos, un análisis computacional general y un análisis computacional específico; y se encuentra dividido en tres secciones.

La primera sección es la introducción, que contiene los objetivos, el alcance y el marco teórico; aquí se detalla la estructura del trabajo y se delimita los temas que abarcará. La subsección del marco teórico realiza una breve revisión de literatura, para luego dar una explicación de conceptos requeridos para que el lector pueda comprender de mejor manera las demás secciones. La segunda sección es la metodología, que contiene la explicación de cómo fue desarrollado cada uno de los análisis y los pasos necesarios para poder reproducir cada análisis. Finalmente, tenemos los resultados, conclusiones y recomendaciones, que contiene información organizada en gráficos y tablas sobre los análisis realizados. Esta sección también tiene conclusiones obtenidas a partir de los hallazgos y recomendaciones para trabajos futuros en esta área.

Para el análisis exploratorio se utilizó los datos del CNE sobre las votaciones del año 2021, y los datos del INEC sobre la encuesta de empleo, desempleo y subempleo. Con los datos se realizó un análisis utilizando el método hipotético-deductivo. Luego se realizó el análisis computacional general acerca del entorno de Kaggle, donde se efectuó el análisis anterior. Esto fue realizado con el fin de poder integrar los temas de sistemas operativos, algoritmos, bases de datos, matemáticas computacionales, teoría de la computación, y cloud computing en el trabajo. De esta manera se analizó la configuración del entorno, sus capacidades, ventajas y desventajas. Finalmente, se procedió a realizar el análisis específico sobre el rendimiento del entorno, mediante el uso del método hipotético-deductivo. Los datos usados para este análisis fueron recolectados durante la ejecución del análisis exploratorio. Estos 3 análisis están expresados en las secciones del trabajo, con los temas divididos según lo correspondiente a cada sección.

1.1. Objetivo general

Demostrar los conocimientos y habilidades adquiridas en las diferentes asignaturas del currículo de la carrera, mediante la realización de un proyecto de análisis exploratorio de datos.

1.2. Objetivos específicos

1. Determinar el comportamiento de variables socioeconómicas y electorales de las elecciones generales del 2021. Definir y comprobar o refutar hipótesis sobre las correlaciones encontradas entre dichas variables, en un documento reproducible.
2. Analizar la infraestructura y organización del entorno utilizado al momento de realizar el objetivo anterior.
3. Analizar el rendimiento de la infraestructura utilizada para la realización del TIC, en un documento reproducible.

1.3. Alcance

Este trabajo toma en cuenta lo siguiente:

Disponibilidad de datos. El trabajo se desarrolló con datos disponibles en la Web pública del CNE [1] y el INEC [2].

Ambiente de reproducibilidad. El trabajo delimita el ambiente computacional (infraestructura tecnológica física y lógica), en el cual se puede reproducir el trabajo (análisis exploratorio y computacional específico) realizado.

Análisis exploratorio. Se realizó un análisis mediante el método hipotético-deductivo para comprender los datos. Para este análisis se utilizó un entorno de R en la nube de Kaggle [3].

Análisis computacional general. Se realizó un análisis cualitativo, de tipo general, sobre el entorno de Kaggle, donde fue realizado el análisis exploratorio. De este entorno se analizó su configuración y estructura.

Análisis computacional específico. Se realizó un análisis del comportamiento que tuvo el entorno al momento de realizar el análisis exploratorio.

1.4. Marco teórico

Revisión sistemática de literatura

Inicialmente se revisó trabajos previos en el área, siendo [4] el primer trabajo revisado; el cual realizó un análisis de la calidad de datos en elecciones alrededor del mundo. El artículo revisó los programas electorales y las opciones para su análisis, como el uso de computadoras, o un análisis humano. Al final concluyó que la calidad de datos disponibles no es suficiente para hacer un análisis concluyente usando computadoras. Adicionalmente,

los autores de [4], dieron acceso público a su conjunto de datos con información acerca de preferencias políticas en cada partido a través de 50 países. Con estos datos se podría hacer un análisis más profundo de las posibles semejanzas y diferencias entre esta información y la de nuestro país. Con todo esto se puede apreciar que los conjuntos de datos públicos, como los usados en el presente trabajo, son una fuente valiosa de información. Por esta razón, se desea poder aportar a esa información creando nuevos conjuntos de datos, y poniéndolos disponibles al público. Los conjuntos de datos y notebooks creados en este trabajo han sido publicados en el sitio web de Kaggle [3], para acceso de cualquier persona que los requiera y pueden ser consultados en los Anexos I y II.

Otro trabajo analizado fue [5], el cual analizó la elección general de Irlanda. Los autores obtuvieron los datos de votaciones electrónicas y determinaron la manera en que los votantes se comportaron con el sistema de elecciones. El sistema de votación que se usó en Irlanda fue mediante el uso de un ranking, pasando votos de los candidatos menos votados al siguiente candidato de preferencia [5]. En este sistema de votación los votantes no asignaron un ranking a aquellos partidos políticos que rechazaban totalmente, asignando ranking a un promedio de tres candidatos. El artículo apreció el comportamiento de los votantes, y los autores pudieron comprender fácilmente qué partidos políticos tienen ideologías similares y comparten votantes. Identificar partidos políticos con ideologías similares es más difícil en el presente trabajo, pero sí es posible identificar si los candidatos recibieron preferencia según el estrato socioeconómico de los votantes. Esto es importante de determinar debido a una percepción general de la población ecuatoriana respecto a las preferencias de candidatos, según el estrato socioeconómico de los votantes.

Finalmente, se revisó [6], donde se pudo apreciar la importancia de los análisis de correlaciones y su uso al comprender mejor el comportamiento de las variables. [6] realizó un análisis de los patrones espacio temporales de voto en Brasil. Además, explicó el pipeline utilizado para los datos, con los pasos seguidos por los autores en el análisis realizado. Los autores encontraron la existencia de una dependencia espacial entre ciudades cercanas, con un análisis que se enfocó en dos partidos políticos que los autores identificaron como los partidos políticos más representativos de Brasil. Los autores de [6] también dejaron disponibles los conjuntos de datos obtenidos de forma pública, tal y como se pretende realizar en el presente trabajo.

Al revisar las investigaciones citadas, se concluyó que ninguno de los investigadores utiliza un método estandarizado para el análisis de los datos. La última investigación incluso crea su propio método para realizar el análisis de datos, al cual lo denomina como un pipeline

sencillo. Por esta razón se tomó la decisión de usar el método hipotético-deductivo, un modelo del método científico, en los análisis que han sido realizados en este componente.

Bases teóricas y definiciones fundamentales

Método hipotético-deductivo (inducción y deducción) [7]: Es un modelo del método científico, siendo la descripción más común que puede ser encontrada en textos científicos. El método consiste en cuatro pasos, los cuales pueden ser repetidos de manera continua para mejorar constantemente el conocimiento. Los pasos descritos en [7] son los siguientes:

1) Observación: Considerar el fenómeno y tratar de entenderlo.
2) Formular una hipótesis para explicar las observaciones (inducción a partir de lo observado): Antes de averiguar más información se intenta establecer una explicación a lo observado, ya sea a partir de la experiencia, o consultando expertos.
3) Deducir posibles consecuencias que deberían ser observables a partir de la hipótesis (deducción): Si asumimos que la hipótesis es verdadera, ¿qué más podríamos observar a partir de ello?
4) Intentar refutar las predicciones realizadas: Se busca información que contradiga estas predicciones. Si no se encuentra información que contradiga las predicciones se puede buscar nuevas predicciones, o realizar nuevas observaciones. Si se encuentra información que contradiga las deducciones se declara la hipótesis como falsa, y se puede crear una nueva.

Podemos observar que una hipótesis nunca llegaría a ser totalmente verdadera, ya que siempre se intentará probar que es falsa. Por otro lado, una hipótesis sí puede ser falsa al demostrar que sus deducciones no se cumplen.

Consejo Nacional Electoral [8]: Conocido por sus siglas CNE, es la máxima entidad electoral en todo el país. Las funciones de este organismo son vigilar, organizar, administrar y certificar las votaciones que se realizan; mediante cálculos, llamamiento a elecciones, publicando resultados, y posesionando a los ganadores.

Elecciones generales del 2021 [9]: Estas elecciones se realizaron en dos vueltas, siendo la segunda vuelta el 11 de abril del 2021. En esta vuelta el ganador para la presidencia fue el binomio Lasso-Borrero con el 52.36% de los votos, mientras que Andrés Arauz perdió la carrera presidencial con el 47.64% de los votos. Para este cálculo sólo se toma en cuenta

los votos válidos, los cuales dejan de lado nulos, blancos y ausentismo (personas que no se acercaron a votar).

Instituto Nacional de Estadísticas y Censos [10]: Conocido por sus siglas INEC, es la institución que se encarga de regular las estadísticas nacionales, las cuales se usan en la política estatal para tomar decisiones.

Encuesta de empleo, desempleo y subempleo (enemdu) [11]: Es conducida por el INEC y está dirigida a los hogares y mide diversos indicadores, como la tasa de desocupación, acceso de agua, entre otros. A partir de estos datos el INEC realiza un cálculo de varias estadísticas de pobreza y mercado laboral, entre los cuales tenemos el estrato socioeconómico, el cual fue usado en el presente trabajo.

Histograma [12]: Es un término usado en estadística, se refiere a un esquema donde una variable se visualiza en forma de barras. Según la frecuencia de los valores la barra será de mayor o menor tamaño.

GPU [13]: También conocido como unidad de procesamiento gráfico, es un tipo especial de procesador diseñado para realizar operaciones de coma flotante de manera paralela. Estas operaciones sirven para procesar gráficos de manera más rápida, lo cual es posible gracias a la gran cantidad de núcleos que posee.

Lenguaje interpretado [14]: Es un lenguaje de programación donde se ejecuta las instrucciones directamente, traduciendo las sentencias en subrutinas que ya han sido compiladas previamente en lenguaje de máquina.

Lenguaje compilado [14]: Es un lenguaje de programación donde antes de su ejecución, un compilador traduce a código máquina todas las sentencias, mediante un compilador.

Endianness [15]: Es un término que designa al esquema que se utiliza para representar datos de más de un byte en una computadora, esencialmente se puede tener los caracteres más representativos al inicio, o al final. Por ejemplo, para representar el número “8A” (en hexadecimal) se puede almacenar primero “8” y luego “A” (little-endian), o primero el “A” y luego “8” (big-endian).

Kernel [16]: Es el software central del sistema operativo, encargado de realizar la gestión de recursos. Decide qué proceso podrá acceder al hardware, y durante cuánto tiempo. En caso de que un proceso pase de su tiempo asignado, se realiza una interrupción, para dejar ingresar a un proceso diferente. Se guarda el estado, ejecutando un cambio de contexto, para volver a ejecutar posteriormente desde el estado anterior.

Kaggle [3]: Es una plataforma para competencias de inteligencia artificial y análisis de datos, ofrece un entorno de notebooks adaptable, con acceso a GPU y grandes cantidades de datos. El modelo de negocios de Kaggle podría definirse como PaaS (Platform as a Service), donde se entrega a los clientes una plataforma para crear recursos. Para participar en una competencia no es necesario realizar pagos, pero para crear una competencia sí, por lo que sus clientes finales son aquellos que crean competencias. El beneficio que reciben los clientes es un software que ha sido refinado a través de la comunidad de competidores que se encuentra en Kaggle.

GNU/Linux [17]: Es un grupo de sistemas operativos compuestos de software libre y código abierto. Surge de la contribución de varios proyectos de código abierto, como son GNU y Linux, donde Linux es el kernel y GNU los componentes del sistema operativo. Al ser de código abierto tiene libertad de uso, estudio, distribución y mejora, conocidas como las cuatro libertades del software libre.

HDFS [18]: Es el sistema distribuido de archivos de Hadoop, que permite almacenar los archivos en un clúster de varios computadores. Gracias a esta característica se puede almacenar grandes cantidades de datos, los cuales pueden ser leídos fácilmente de forma secuencial. Es un sistema escalable, y proporciona redundancia, dándole la posibilidad de acceso paralelo y tolerancia a fallos.

Vmstat [19]: es un programa que puede ser ejecutado en GNU/Linux y permite obtener datos sobre procesos, paginación, E/S y memoria, del computador.

Lenguaje R [20]: Es un lenguaje de programación interpretado, enfocado al análisis estadístico y muy utilizado en investigación científica, es uno de los componentes distribuidos como parte de GNU. Este lenguaje otorga la posibilidad de desarrollar bibliotecas en C, C++ o Fortran, las cuales son cargadas dinámicamente, y permiten un desempeño más rápido al ser lenguajes compilados a diferencia de R.

Gráfico de cuantiles (QQ) [21]: Es un gráfico usado para visualizar las diferencias en la distribución de probabilidad. Se calcula entre una población de donde se ha obtenido una muestra y una distribución usada para comparación, que puede o no ser teórica. Para determinar la distribución se realiza un cálculo para cada uno de los datos de forma individual, y se le asigna al cuantil al que pertenece.

Análisis de componentes principales (PCA) [22]: Es una técnica que describe los datos a través de nuevas variables o componentes, los cuales se ordenan según cómo describen la varianza de los datos originales. Esta técnica busca la mejor representación de los datos,

con la menor cantidad posible de componentes. Para realizar estos cálculos se utiliza matrices de covarianza, siendo las operaciones entre matrices los cálculos más predominantes.

Correlación de Pearson [23]: Es un valor que mide la dependencia lineal entre dos variables, es independiente de la escala usada en las medidas y se utiliza para medir el grado de relación lineal entre dos variables. El valor obtenido varía entre -1 y 1, lo cual nos indica el sentido de la relación, con 1 siendo una correlación positiva perfecta, -1 siendo una correlación negativa perfecta, y 0 indicando no correlación en absoluto. En los casos de correlación positiva si una variable aumenta la otra también, mientras que cuando la correlación es negativa, si una variable aumenta, la otra disminuye.

K-means [24]: Conocido como k-medias en español, es un método que pretende seccionar un conjunto en k grupos, cuyo valor medio o centroide es el más cercano entre sí. Este algoritmo se calcula obteniendo el valor medio de cada grupo aleatorio definido inicialmente, para luego redistribuir los grupos según el centroide más cercano, y volver a realizar el cálculo. Para finalizar las iteraciones se define una tolerancia de variación entre los centroides obtenidos entre un cálculo y el siguiente.

2. METODOLOGÍA

Para el desarrollo del componente se utilizó el método hipotético-deductivo, con la lectura de datos en lugar de la experimentación en el análisis exploratorio, y experimentación en el análisis computacional específico. Se comenzó explorando los datos, para luego plantear varias hipótesis, de las cuales se analizó las consecuencias, para luego intentar demostrar que las hipótesis sean falsas. Esta metodología fue tomada de [7], y los temas han sido tratados de acuerdo a las componentes del trabajo definidas en el alcance.

- ***Enfoque:*** El enfoque del componente es mixto, ya que tuvo un enfoque cuantitativo tanto en el análisis exploratorio, como en el análisis computacional específico; pero tuvo un enfoque cualitativo en el análisis computacional general.
- ***Tipo de trabajo:*** Es un trabajo exploratorio, descriptivo y experimental; ya que a partir de los datos disponibles se realizó el análisis completo de cada paso del método utilizado. En la sección de análisis computacional específico se realizó experimentación.
- ***Técnica de recolección de información:*** Se utilizó el análisis documental, ya que la información viene de fuentes oficiales, (INEC y CNE). A partir de esta información se realizó el análisis exploratorio, del cual se obtuvo datos experimentales con la herramienta vmstat para el análisis computacional específico.
- ***Técnica, basada en estadística, para el análisis de información:*** Para el análisis de la información se utilizó histogramas, gráficos de cuantiles, tablas, análisis de componentes principales, correlación de Pearson y clústeres.

En este componente se realizó tres análisis distintos: exploratorio, computacional general y computacional específico. El método hipotético-deductivo fue utilizado en los análisis exploratorio y computacional específico, dado que el análisis computacional general es de tipo descriptivo y analiza la infraestructura del entorno computacional.

El análisis exploratorio fue realizado para comprender los datos de las elecciones generales del 2021. A partir de estos datos se extrajo hipótesis que se enfocaron en comprender el comportamiento de las personas al momento de realizar su voto. El análisis computacional general fue ejecutado debido a que se deseaba comprender la configuración del entorno utilizado. Finalmente, el análisis computacional específico fue realizado con la intención de verificar cómo se desempeña el entorno computacional con la ejecución del análisis exploratorio.

2.1. Análisis exploratorio

Los datos usados para el análisis fueron descargados de dos sitios web, el sitio web del CNE [1] y el sitio web del INEC [2]. En el sitio web del CNE se obtuvieron los datos de las elecciones generales del 2021, tanto de segunda vuelta para presidente, como del registro electoral. En el sitio web del INEC se obtuvieron los datos de la encuesta de empleo, desempleo y subempleo, de marzo del 2021. Esto debido a que fue la última encuesta realizada antes de la segunda vuelta de las elecciones generales. Los datos de ambos sitios web fueron subidos al entorno en la nube de Kaggle, donde se realizó el análisis exploratorio. La carga se realizó a través de la interfaz de usuario disponible en Kaggle, creando un conjunto de datos y un documento reproducible, el cual puede ser encontrado en el Anexo I. Para este análisis se utilizó el método hipotético-deductivo, tal y como está descrito en el marco teórico del presente trabajo, y que fue tomado de [7].

Método hipotético-deductivo:

a) Observación de los datos.

Los datos están organizados en una tabla, con nombres en cada columna, donde se comenzó cambiando los nombres de las columnas por unos nombres más entendibles. Luego se analizó por separado los datos de cada candidato para determinar el ganador de la elección, siendo éste el candidato 1021; también se sumó los sufragantes que se tuvo en las elecciones. Con estos datos se hizo una comparación entre el total de votos (válidos, más blancos y nulos) con el total de sufragantes. Aquí se obtuvo que el total de votos no coincide con el número total de sufragantes, debido a que se encontraron 10829823 sufragantes y 10828723 votos totales. Por otro lado, también se revisó los datos del registro electoral, para poder verificar el ausentismo al comparar con los datos de los sufragantes; el ausentismo fue del 17%.

Gracias a los datos de la encuesta de empleo, desempleo y subempleo se observó los estratos socioeconómicos de cada provincia. En esta encuesta se obtiene los estratos socioeconómicos de bajo, medio y alto, los cuales vienen con números: 1 para bajo, 2 para medio y 3 para alto. Los datos se organizaron en una tabla, para poder destacar las provincias según su promedio de estrato socioeconómico. Las provincias que destacaron fueron Morona Santiago, con el estrato socioeconómico más alto en promedio (2.161), y Galápagos con el estrato socioeconómico promedio más bajo (1.795). Gracias a la obtención de estos datos se pudo analizar correlaciones entre el estrato socioeconómico de una provincia y el número de votos de los candidatos en la misma.

b) Creación de hipótesis.

A partir de las observaciones se creó dos hipótesis: una basada en la diferencia entre el total de votos y sufragantes, y otra basada en una percepción común en redes sociales respecto a los votantes de cada candidato.

I) La diferencia entre el total de votos y el total de sufragantes se debió a errores efectuados por los miembros de las juntas receptoras de voto.

II) Los votantes tuvieron preferencia por algún candidato en específico según el estrato socioeconómico al que pertenecían.

c) Deducción de las consecuencias de la hipótesis.

Después de analizar las hipótesis, se determinó las posibles consecuencias que se podrían observar, si las hipótesis fueran verdaderas.

I) Consecuencias de la primera hipótesis.

- El número de sufragantes, y la cantidad de votos que no coincidieron, deberían tener una distribución estadística similar. Si la diferencia se debió a errores comunes, estos errores serían más frecuentes mientras más sufragantes existan, se debería de observar alguna correlación entre ambas variables.
- El porcentaje de votos que no coincidieron con respecto a la cantidad de sufragantes debería ser bajo, y similar en todas las zonas analizadas.
- Las parroquias que se comportaron de forma anómala deberían hacerlo por razones ajenas a la cantidad de votos que no coinciden.

II) Consecuencias de la segunda hipótesis.

- En caso de que los votantes tuvieran preferencia por un candidato según su estrato socioeconómico, se apreciaría una correlación entre los votos del candidato y el estrato socioeconómico promedio de cada provincia.

d) Comprobar o refutar las deducciones.

A continuación, se analizó los datos para intentar refutar las deducciones realizadas. Primero se organizó la información para verificar el comportamiento de las variables (número de sufragantes, diferencia con votos válidos, estrato socioeconómico, etc.). Una vez que la información estuvo organizada se analizó las variables para poder determinar el comportamiento que poseen.

I) Procedimiento usado para la primera hipótesis.

Para esta hipótesis los datos de votaciones fueron agregados por parroquia, debido a que los datos del registro electoral vienen agregados de la misma manera. Luego se procedió a analizar la distribución de los datos mediante el uso de histogramas. La información de estos histogramas fue de los sufragantes, diferencia de votos y electores; y muestran la frecuencia de valores que aparecen en cada una de las columnas. En este tipo de gráficos, un valor más alto significa que existen más datos con esos valores.

En el siguiente paso se analizó los datos mediante el gráfico QQ comparando la distribución de los datos con una distribución normal. También se realizó un análisis de correlaciones entre las tres variables a disposición, intentando comprender si la variación de una de ellas sucede de manera conjunta con la variación de otra. Adicionalmente, se realizó un análisis de componentes principales, donde se analizó la varianza de los datos con cada reducción de componentes. Si la varianza es muy similar a la varianza original, entonces se asume que el número de componentes es aceptable. El objetivo es conseguir la menor cantidad posible de componentes. Al haber reducido el número de variables se elimina las variables redundantes, lo que permite distinguir mejor los datos.

A continuación, se buscó anomalías en los datos mediante la creación de un clúster para observar aquellos datos que se encuentran más alejados del centro; se utilizó un único clúster para determinar de manera general los datos anómalos con respecto al resto. Luego, se obtuvieron los 10 puntos más alejados del centro, para intentar comprender por qué se comportan de forma diferente. Finalmente, se realizó tres histogramas mostrando las 30 parroquias que tienen un número más alto de “diferencia de votos”, “número de sufragantes” y “número de electores”. El primer histograma mostró la cantidad de votos diferentes en cada parroquia; el segundo histograma mostró los mismos datos como porcentaje de la cantidad de sufragantes; y el tercero mostró los datos como porcentaje de la cantidad de electores. De esta manera se pudo encontrar las parroquias con un mayor número de diferencia de votos, y determinar si éstas también destacan al ver los datos como porcentaje.

II) Procedimiento usado para la segunda hipótesis.

Para analizar esta hipótesis primero se agrupó los datos del INEC, y se adjuntó esos datos a los obtenidos del CNE. Estos datos permitieron visualizar de manera más clara el comportamiento de las distintas columnas, y sus relaciones. Las variables que fueron agrupadas son: estrato socioeconómico, cantidad de sufragantes y electores, diferencia en

votos, votos de cada candidato, blancos, y nulos. Los votos fueron calculados a modo de porcentaje respecto al número de sufragantes.

Después de la agrupación, se realizó histogramas de frecuencia para distribución de los datos. También se realizó los gráficos QQ, para determinar si la distribución de las variables se ajusta a una distribución normal. Además, se realizó un análisis de correlaciones, para determinar si existen variables que se mueven de manera conjunta en los datos analizados. Por último, se realizó un análisis de componentes principales para determinar si se tiene componentes redundantes y cuántos serían.

2.2. Análisis computacional general

Para determinar el tipo de entorno utilizado por Kaggle se recurrió al uso del paquete "benchmarkme", el cual permite acceder a la información del sistema. Con esto se pudo obtener el sistema operativo, la arquitectura, el procesador, la versión de R, la memoria RAM y la información de los paquetes de álgebra lineal que fueron instalados.

Sistema Operativo

Kaggle usa el sistema operativo GNU/Linux, el cual podría brindar ventajas para nuestro análisis exploratorio, como el hecho de tener las cuatro libertades del software libre. Adicionalmente, la arquitectura usada por Kaggle es de little-endian, lo que significa que tiene pequeñas ventajas en transformaciones entre distintos tipos de datos, lectura de variables y realización de pequeñas operaciones matemáticas. La principal ventaja de big-endian, en cambio, sería de poder utilizar un formato similar al que se usa para imprimir, y que además es el formato más utilizado para la transmisión de datos a través de los protocolos de internet. En este caso little-endian es más ventajoso para realizar los análisis.

Infraestructura física

El hardware utilizado por Kaggle se obtiene utilizando las funciones de get_ram y get_cpu. Así se observa las diferencias sutiles entre el entorno con GPU y el entorno de CPU, como el cambio de procesador o la memoria RAM ligeramente reducida. Para el caso del entorno con CPU se tiene 18.9 GB de memoria RAM y el procesador puede ser un Intel® Xeon® de 2.20GHz o un AMD EPYC 7B12 2.25GHz, ambos de 4 núcleos. En el caso del entorno con GPU, la memoria RAM es un poco menor, de 16.8 GB, y el procesador es un Intel® Xeon® de 2.00GHz de dos núcleos. Adicionalmente, se tiene el GPU del entorno, el cual es Tesla P100-PCIE-16GB.

En el caso del procesador para el entorno de CPU, se tiene dos opciones de procesamiento, ambos procesadores tienen cuatro núcleos y una velocidad de

procesamiento similar. Por esto se puede asumir que están en el mismo rango, el cual es superior a la versión disponible en el entorno de GPU, con solo 2.00GHz y dos núcleos. Esto sumado a la diferencia en memoria RAM indica que los entornos están enfocados a diferentes casos de uso. En el caso de que se use un GPU, se espera que los cálculos realizados por los procesadores serán menores a los que se realizarán en un entorno de CPU, ya que algunos cálculos serán enviados al GPU. De igual manera la reducida memoria RAM no afecta mucho al entorno de GPU, porque el GPU tiene su propia memoria RAM dedicada, con la cual accede de manera más directa a los datos para procesamiento. Finalmente, el GPU Tesla P100-PCIE-16GB nos permite realizar cálculos paralelos mediante el uso de Cuda, acelerando cálculos como las operaciones entre matrices.

Posibilidades de computación paralela y distribuida

Entre los algoritmos utilizados existen cuatro algoritmos con la posibilidad de computación paralela, o distribuida, los cuales son PCA, K-means y el gráfico de cuantiles. Estos algoritmos realizan múltiples cálculos entre todos los datos, y pueden ser realizados en forma simultánea ya que muchos cálculos no dependen de los resultados de cálculos previos. En el caso de PCA, las operaciones de matrices pueden ser paralelizadas fácilmente [22], K-means permite paralelizar al separar los datos para los cálculos [24], ya que el promedio puede ser calculado por partes. Para el gráfico QQ se puede calcular igualmente a cuál cuantil pertenece al separar los datos, ya que el cálculo es individual [21].

Base de Datos

En Kaggle los datos son cargados a HDFS, lo cual se puede observar al leer la descripción de la consola mientras el entorno de notebooks es iniciado. Este sistema permite leer y almacenar archivos de gran tamaño, y leerlos de forma paralela [18]. Lo anterior es útil en Kaggle, ya que los usuarios pueden subir archivos individuales, carpetas, archivos comprimidos, entre otros; con un límite de 100GB por archivo [3]. Con archivos tan grandes es imperativo usar un sistema de almacenamiento que sea fácilmente escalable y permita lecturas rápidas de los ficheros.

Seguridad de los datos

Los datos contenidos en las bases del CNE [1] presentan información que ha sido anonimizada, por lo que no se ha encontrado fallas de seguridad en esos datos. En el caso de los datos de elecciones se puede encontrar la información de diferentes niveles, llegando al nivel de parroquia, los cuales no permitirían identificar a las personas que

efectuaron su derecho al voto. En los datos de padrón electoral, se puede igual ver los datos hasta nivel de parroquia, sin tener información más específica de las personas. En el caso de la información obtenida del INEC [11], se tiene que las personas fueron anonimizadas, al asignarles un código único imposible de rastrear.

Lenguaje R

El lenguaje R al ser un lenguaje interpretado permite realizar de manera más sencilla los análisis de datos. Esto es debido a que se puede ejecutar las líneas de código de manera independiente, almacenando la información de variables en memoria para poder accederlas después. Sin embargo, como se expresó en el marco teórico, un lenguaje interpretado es más lento, por lo que R usa paquetes compilados en otro lenguaje [20]. De esta manera el usuario de R puede obtener un mejor desempeño, ya que los lenguajes compilados son más eficientes. Sin embargo, esto limita en cierta manera el uso de R a los paquetes ya existentes, debido a que el desempeño de los programas desarrollados en R directamente nunca podrá igualarse al desempeño de los paquetes previamente compilados.

2.3. Análisis computacional específico

Este análisis se realizó para comprender mejor el comportamiento de los algoritmos usados en el análisis exploratorio previo. Adicionalmente, con este análisis se pretende obtener un mayor entendimiento de cómo se comporta la infraestructura, física y lógica, con el código ejecutado. De igual manera que con el análisis exploratorio, este análisis utilizó el método hipotético-deductivo definido en [7].

Método hipotético-deductivo:

a) Observación de los datos.

La observación inicial en este caso fue realizada mientras se ejecutaba el análisis exploratorio. Con base en el funcionamiento determinado en el marco teórico de los algoritmos utilizados, se esperaba que ciertas celdas de código tengan una ejecución más intensiva en recursos. Los algoritmos que se espera sean los más que más recursos demanden son: K-means, correlación de Pearson, análisis de componentes principales, comparación entre cuantiles y carga de datos.

b) Creación de hipótesis.

A partir de lo observado, se realizó una predicción de cuáles celdas tendrían mayor consumo de recursos al ejecutarse. En primer lugar, se esperó que las celdas 3, 6 y 8

transfieran bastantes datos, ya que son las celdas que se encargaron de cargar los datos. Las celdas 10 y 20 también se esperó que fueran intensivas para el entorno, debido a que se encargaron de realizar los análisis de componentes principales, correlaciones entre variables y comparación de cuantiles. Finalmente se esperó que la celda 12 también fuera intensiva para el entorno, debido a que ahí se realizó la creación del clúster con el algoritmo K-means.

Con estas observaciones se definió la siguiente hipótesis:

- Las celdas 3, 6, 8, 10, 20 y 12 consumieron más recursos computacionales que las demás al momento de ejecutarse el análisis exploratorio.

c) Deducción de las consecuencias de la hipótesis.

Las consecuencias en caso de que la hipótesis sea verdadera son las siguientes:

- Los recursos computacionales consumidos mostrarán valores más altos en las celdas 3, 6, 8, 10, 20 y 12.

d) Procedimiento para intentar refutar las deducciones.

Para comprobar o refutar nuestra hipótesis se diseñó un experimento sencillo para reunir datos, los cuales se obtuvieron del entorno virtual durante la ejecución del análisis exploratorio, con el uso de la herramienta vmstat. Esta herramienta fue iniciada mediante un comando del sistema previo al análisis exploratorio enviado desde R con el uso de la función “system”. Se ejecutó vmstat durante 150 segundos, a un intervalo de un segundo y con la información de tiempo. Además, se obtuvo una salida de datos amplia y con el encabezado mostrado una única vez, para luego imprimir el resultado en el archivo vmstat.dat.

Para poder comparar la ejecución de las celdas con la información de vmstat, se imprimió también una marca de tiempo al inicio de cada una de las celdas. Así se pudo visualizar de mejor manera el comportamiento del entorno virtual, y relacionarlo con los comandos que estaban siendo ejecutados en ese momento. La ejecución del análisis exploratorio se realizó 10 veces para cada entorno (20 en total), intentando obtener datos con mayor estabilidad. Finalmente, estos datos fueron recopilados descargando el resultado, en el caso de los datos de vmstat. Mientras que, para los datos de la marca de tiempo de cada celda, se realizó una inspección manual.

Una vez que los datos fueron recopilados, se los subió a un fichero en Kaggle, donde se tiene los datos obtenidos de vmstat y también los datos de la marca de tiempo en cada una de las celdas. Estos datos fueron organizados en carpetas para luego ser leídos en R,

donde se separó la información, y se renombró a cada dato. Con el uso de un bucle, se obtuvo la información de cada una de las ejecuciones, y se obtuvo promedios de cada celda sobre las métricas de vmstat. Se diferenció entre GPU y CPU para intentar establecer si hubo alguna mejora en alguno de los entornos, la cual debería haber sido implementada de forma automática por las librerías, ya que no se hizo ninguna modificación manual. Este análisis se ejecutó en un entorno virtual de R similar al usado para el análisis exploratorio, y puede ser encontrado en el Anexo II. Para visualizar de mejor manera los datos, se recurrió al uso de gráficos de líneas, donde se pudo ver cómo evolucionan estas métricas a medida que avanza el análisis exploratorio efectuado.

Las medidas entregadas por vmstat son sobre procesos, memoria, swap, bloques, sistema y tiempo del CPU. Adicionalmente, se recopiló la duración de cada celda a partir de los datos de la marca de tiempo que se tiene en cada una. Es preciso aclarar que para los datos de duración se tiene precisión en segundos, de modo que los datos reportados están redondeados al inmediato superior.

3. RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1. Resultados

Análisis exploratorio

I) Primera hipótesis:

Los histogramas de las variables “diferencia de votos”, “electores” y sufragantes pueden verse en la Figura 3.1, donde se observó que la distribución de las tres columnas no fue una distribución normal. Además, las columnas tuvieron una distribución muy similar entre sí, más similar entre “electores” y “sufragantes”.

En el gráfico entre cuantiles, o gráfico QQ, de la Figura 3.2 se puede verificar lo visualizado anteriormente sobre la distribución de los datos. La distribución no fue normal, pero hubo una gran similitud entre las columnas.

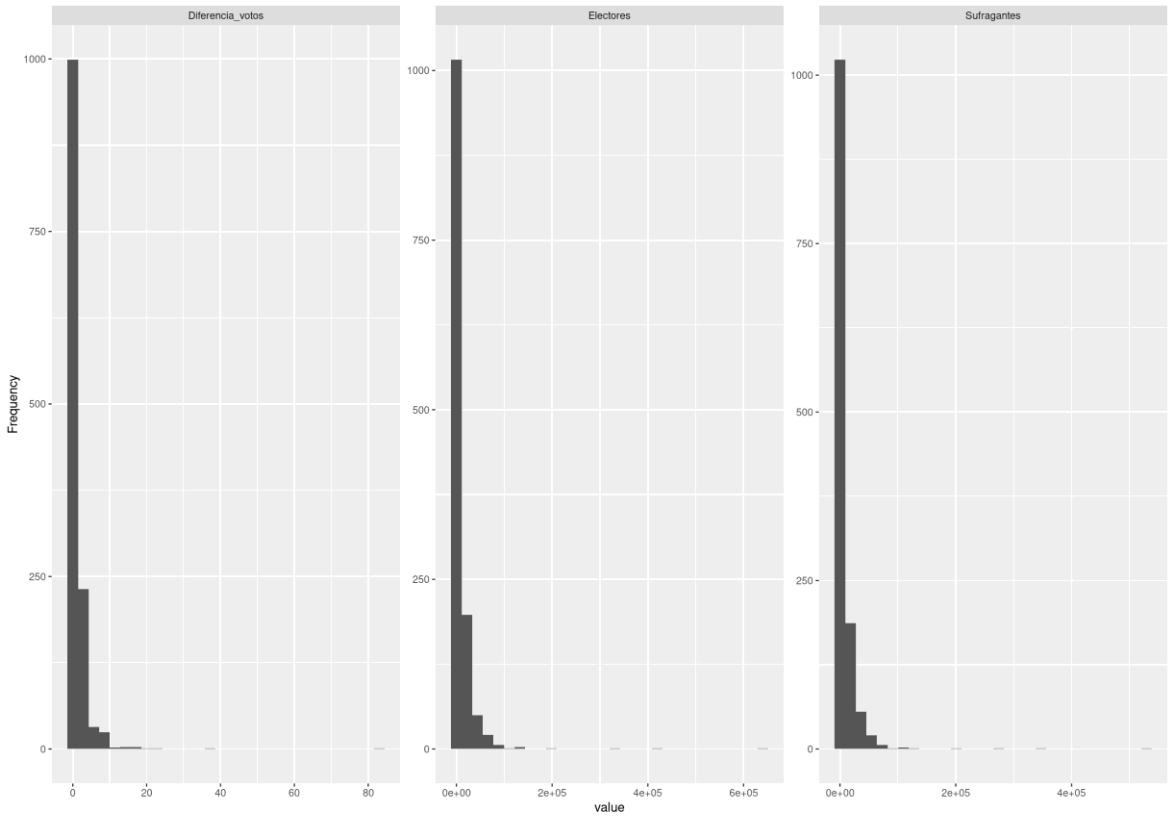

Figura 3.1. Histogramas de las variables de votaciones.

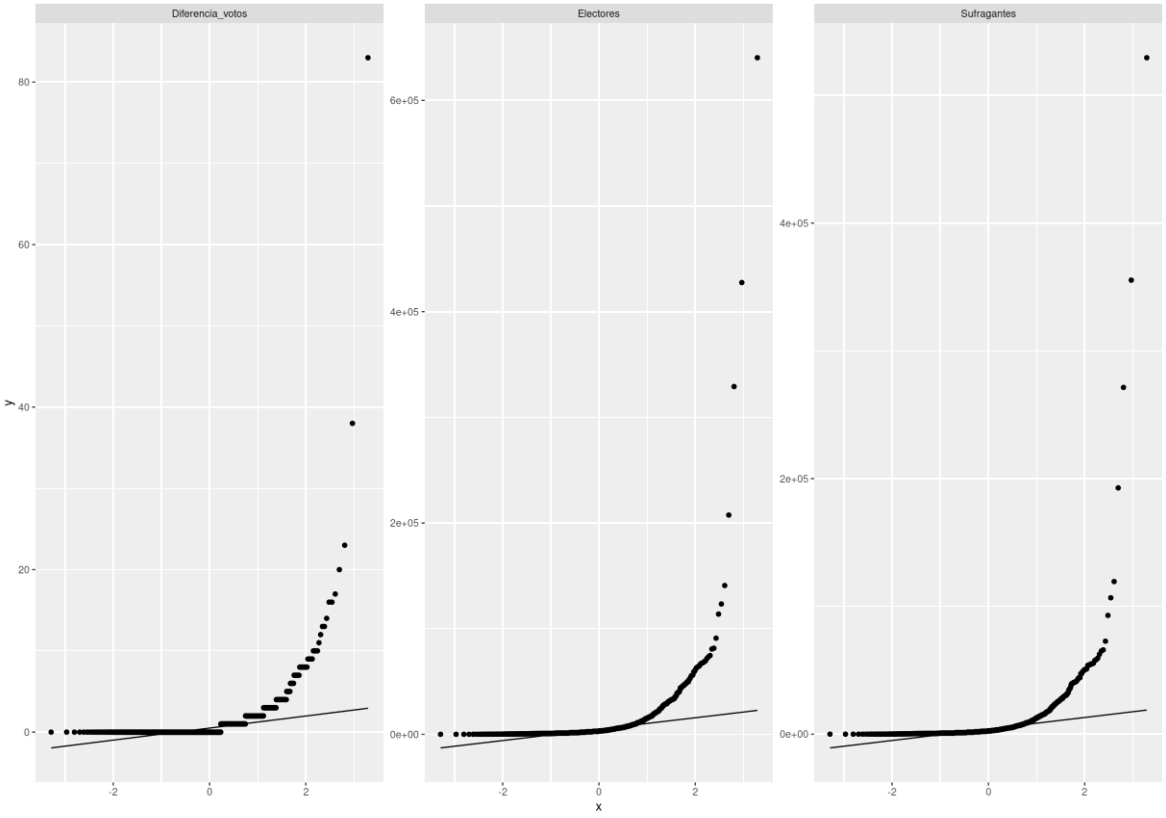

Figura 3.2. Gráfico QQ de los votos, sufragantes y electores.

En el análisis de correlaciones de la Figura 3.3 se observa que las variables de “electores” y “sufragantes” tuvieron el máximo valor de correlación (siendo 1). De modo que, si aumentaba el número de electores, también aumentaba el número de sufragantes, y viceversa. También se pudo apreciar que la columna de “diferencia de votos” tiene una correlación muy fuerte con “electores” y “sufragantes” (0.86). Así, si la diferencia de votos es más alta, también “sufragantes” y “electores” serán más altos.

Por otro lado, en el análisis de componentes principales, el 94% de toda la varianza de los datos pudo ser expresada por un único componente; esto se debió a que las tres variables eran muy similares entre sí. Además, se puede observar en la Figura 3.4 que las variables de “electores” y “sufragantes” tuvieron una importancia mayor en el componente seleccionado.

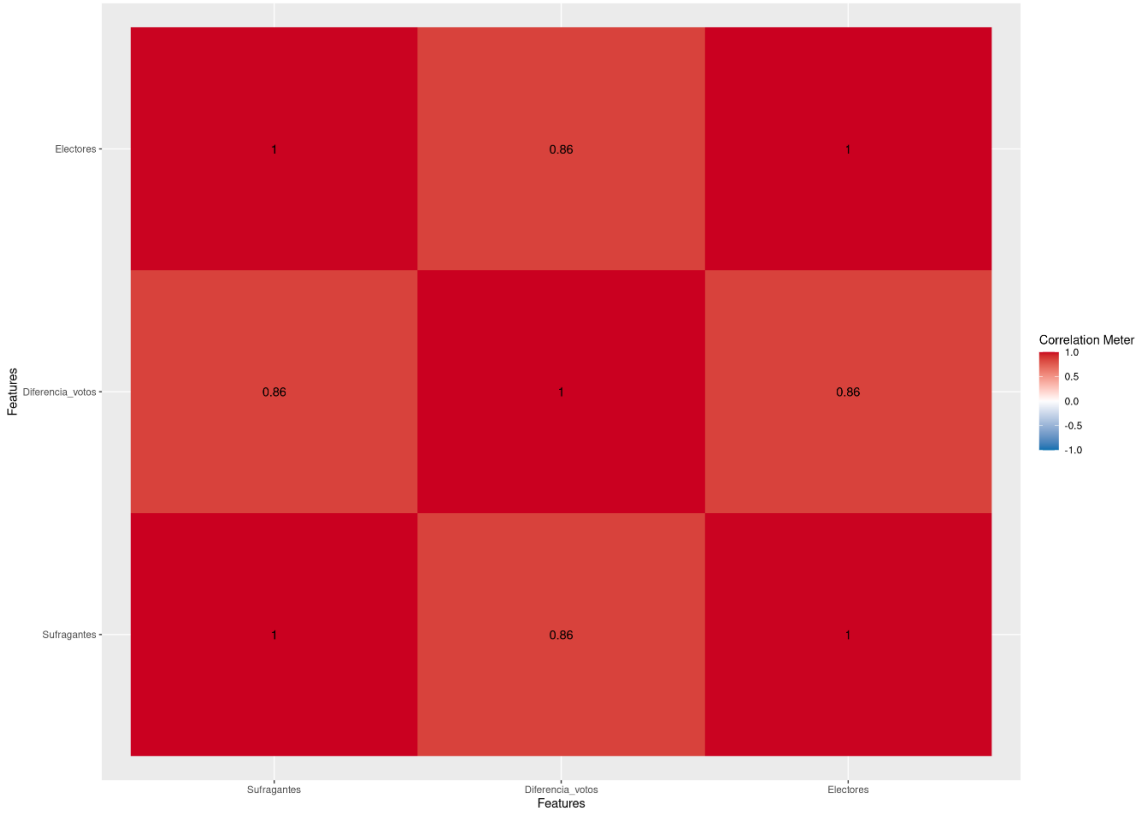

Figura 3.3. Correlación entre variables de votos.

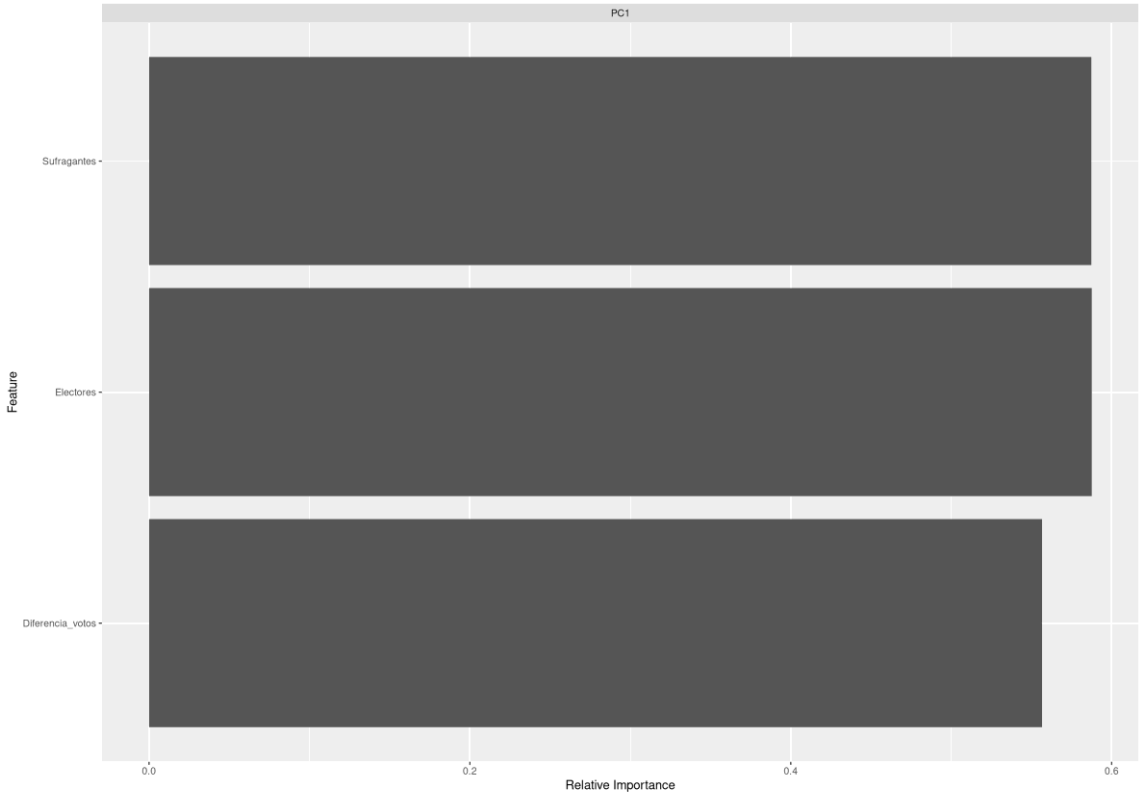

Figura 3.4. Importancia de componentes en los votos.

El resultado de la creación de un clúster con el algoritmo K-means se observa en la Figura 3.5, donde se puede ver que existen varios puntos alejados del clúster, siendo considerados anómalos los más alejados del centro. Se utilizó un único clúster, debido a que se pretende conseguir datos anómalos de manera general respecto a los demás datos.

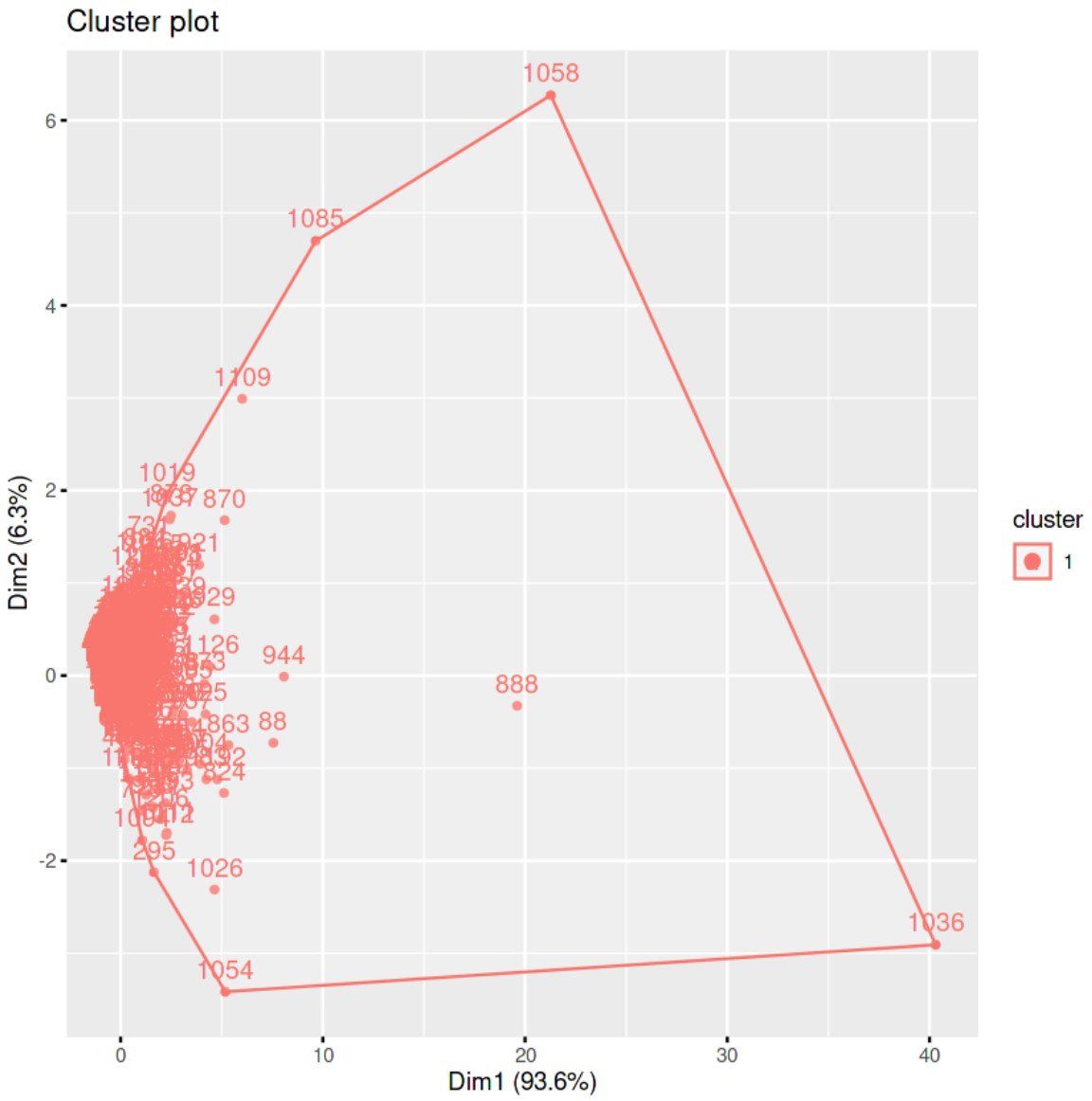

Figura 3.5. Clúster creado para mostrar anomalías.

En la Tabla 3.1 se visualiza los 10 puntos más alejados del centro. En estas parroquias la diferencia de votos fue grande, pero también la cantidad de electores y sufragantes. Adicionalmente, al comparar con la Tabla 3.2, donde se puede observar las parroquias con mayor número de electores, se visualiza que básicamente son las mismas parroquias.

Tabla 3.1. Datos de las parroquias identificadas como anomalías.

Provincia	Cantón	Parroquia	Sufragantes	Diferencia de votos	Electores	Distancia del centro del clúster
GUAYAS	GUAYAQUIL	TARQUI	529448	83	640404	816.49013
GUAYAS	GUAYAQUIL	XIMENA	355347	23	427610	245.94124
GUAYAS	GUAYAQUIL	FEBRES CORDERO	271391	38	329153	192.29729
GUAYAS	GUAYAQUIL	PASCUALES	192764	6	207525	57.74253
GUAYAS	DURAN	ELOY ALFARO /DURAN	117012	4	142010	22.53349
GUAYAS	MILAGRO	MILAGRO	119460	16	140850	32.61235
PICHINCHA	QUITO	CALDERON	106719	17	123371	28.80349
PICHINCHA	QUITO	CHILLOGALLO	92957	6	113792	14.64553
PICHINCHA	QUITO	COTOCOLLAO	72747	13	90990	14.51999
LOS RIOS	VINCES	VINCES	49322	20	55414	19.23598

Tabla 3.2. Datos de las parroquias con mayor número de electores

Provincia	Cantón	Parroquia	Sufragantes	Diferencia de votos	Electores	Distancia del centro del clúster
GUAYAS	GUAYAQUIL	TARQUI	529448	83	640404	816.49013
GUAYAS	GUAYAQUIL	XIMENA	355347	23	427610	245.94124
GUAYAS	GUAYAQUIL	FEBRES CORDERO	271391	38	329153	192.29729
GUAYAS	GUAYAQUIL	PASCUALES	192764	6	207525	57.74253
GUAYAS	DURAN	ELOY ALFARO /DURAN	117012	4	142010	22.53349
GUAYAS	MILAGRO	MILAGRO	119460	16	140850	32.61235
PICHINCHA	QUITO	CALDERON	106719	17	123371	28.80349
PICHINCHA	QUITO	CHILLOGALLO	92957	6	113792	14.64553
GUAYAS	GUAYAQUIL	LETAMENDI	74907	8	95360	10.94638
PICHINCHA	QUITO	COTOCOLLAO	72747	13	90990	14.51999

Finalmente, en los histogramas de las Figuras 3.6, 3.7 y 3.8 se observa que, en cantidad de diferencia de votos, destacó la parroquia de Tarqui (Guayaquil), pero en porcentaje destacó en ambos casos la parroquia de Amazonas (Morona Santiago). Los gráficos muestran que la cantidad de votos diferentes fue mínima al compararla con la cantidad de sufragantes, o con la cantidad de electores, con un porcentaje menor al 1%.

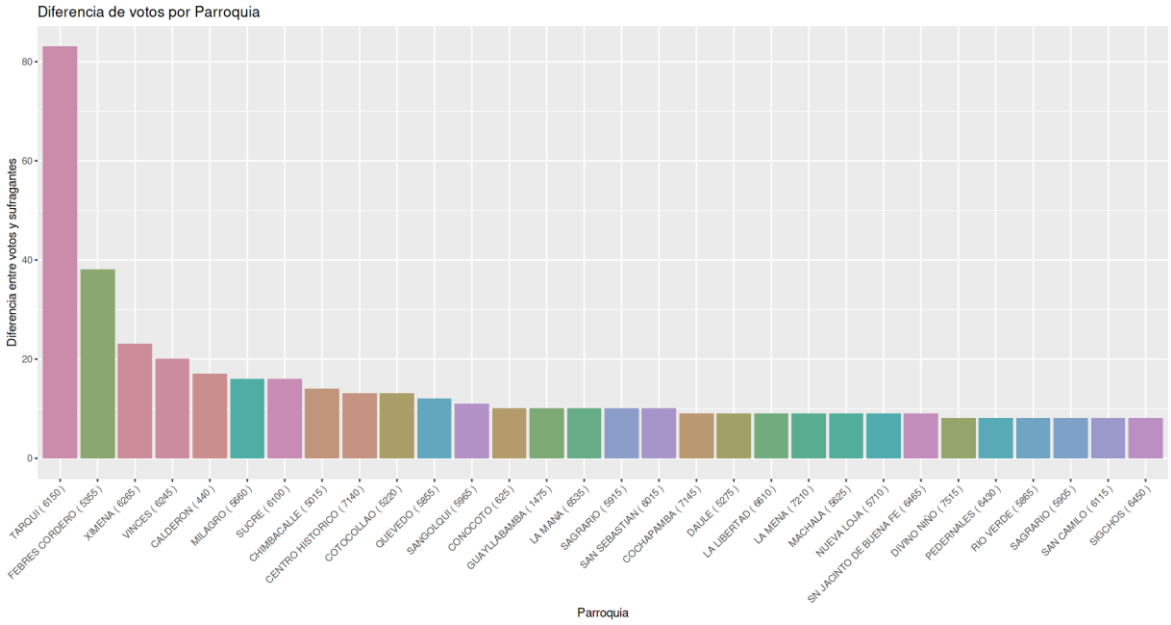

Figura 3.6. Diferencia de votos por parroquia.

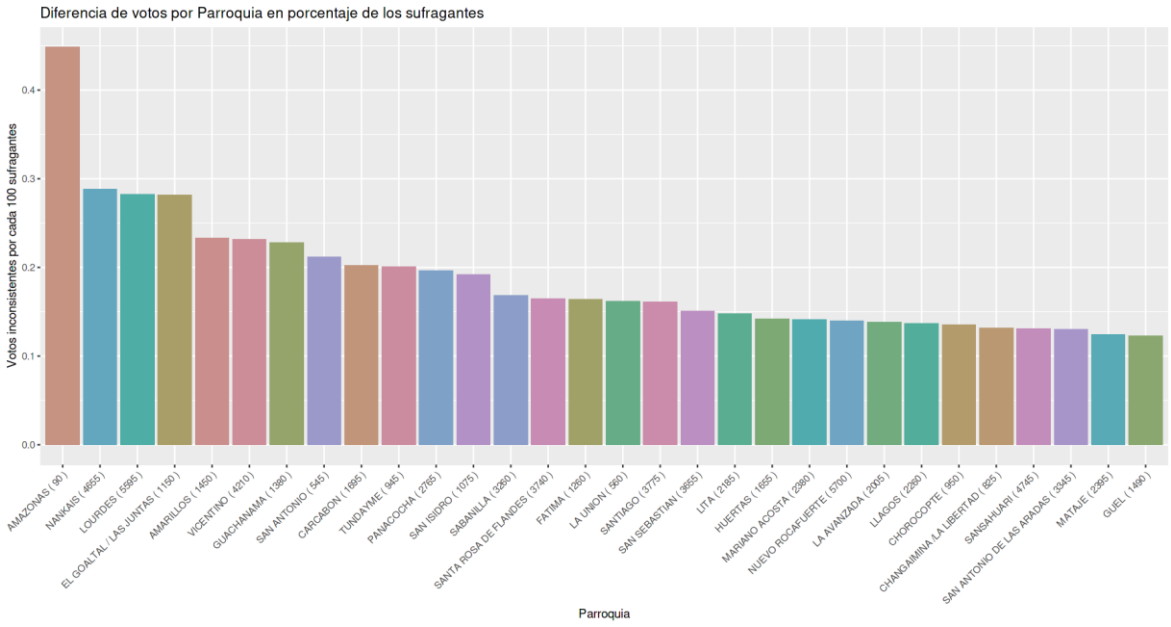

Figura 3.7. Diferencia de votos por parroquia en porcentaje de sufragantes.

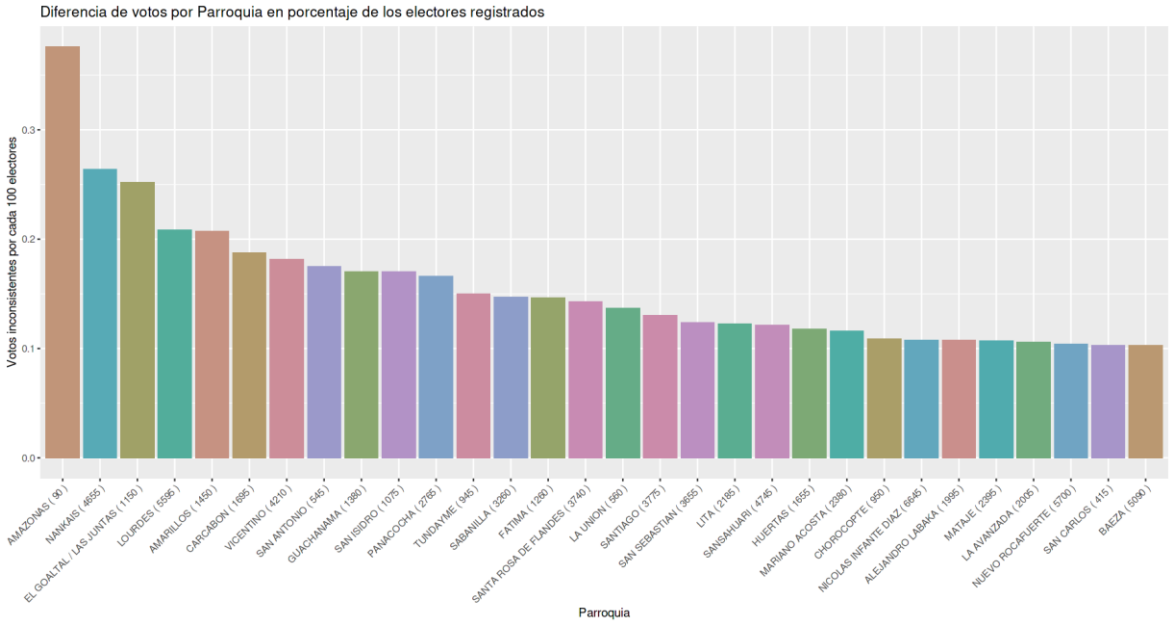

Figura 3.8. Diferencia de votos por parroquia en porcentaje de electores.

II) Segunda hipótesis:

En los histogramas de la Figura 3.9, no se observa un comportamiento específico que haya sido similar entre las variables. Por otro lado, se puede observar que existieron variables con distribución muy parecida a la normal, como fueron los votos, blancos y nulos, y el estrato socioeconómico.

En la Figura 3.10 se observa los gráficos entre cuantiles, donde las variables de “votos blancos”, “votos nulos”, y “estrato socioeconómico” tuvieron un comportamiento similar al de una distribución normal, sin mucha similitud entre ellas. En el análisis de correlación, visto en la Figura 3.11, se puede notar unas correlaciones importantes. Inicialmente, se aprecia que la correlación entre electores, sufragantes y diferencia de votos sigue presente. Sin embargo, esa no fue la única correlación existente; también se observa que la cantidad de votos nulos se correlacionó con muchas variables. Existió una correlación negativa entre “sufragantes”, “electores” y “diferencia de votos”, con “votos nulos”; mientras más electores tuvo una provincia, un menor porcentaje de ellos votó nulo. Adicionalmente, los votos nulos se relacionaron con el estrato socioeconómico de manera positiva; mientras más votos nulos existieron en una provincia, más alto fue el estrato socioeconómico de esa provincia. Los votos nulos no tuvieron ningún tipo de correlación con los votos de Guillermo Lasso, pero sí tuvieron una correlación negativa con los votos para Andrés Arauz. Finalmente, hubo una correlación positiva entre los votos nulos y los votos blancos, mientras más votos nulos había, más votos blancos también estaban registrados en esa provincia.

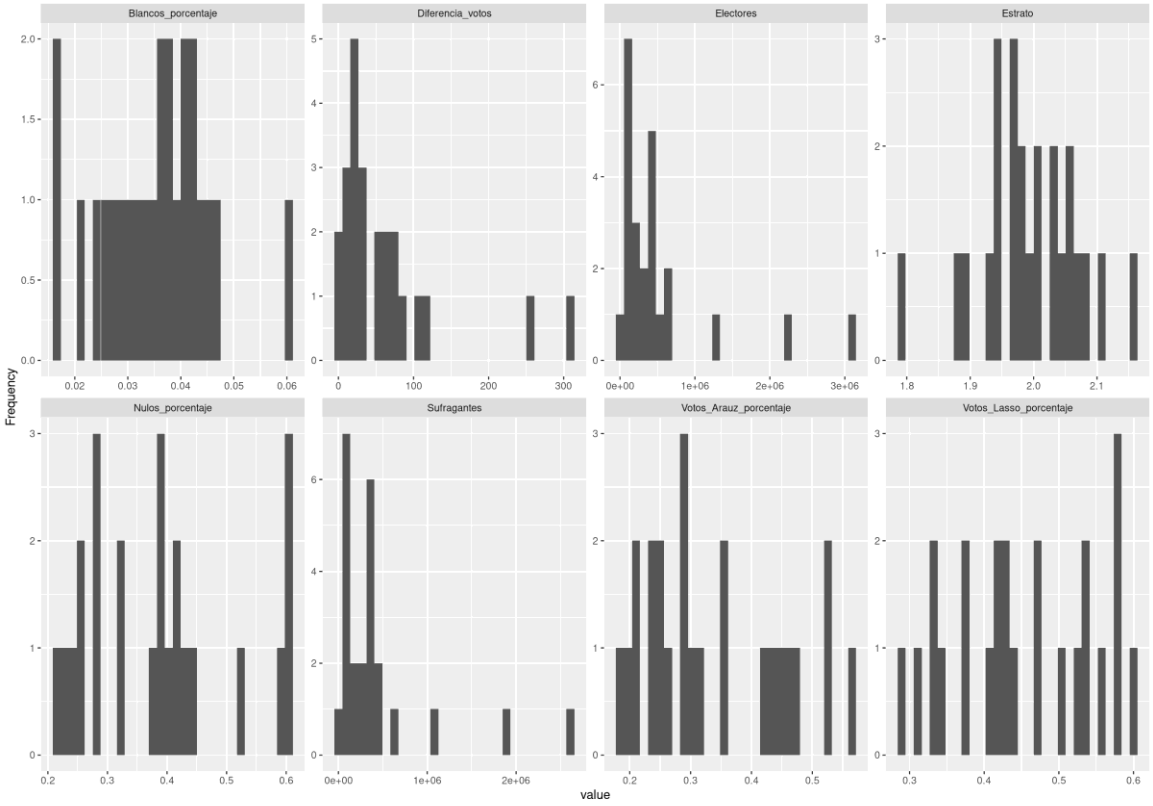

Figura 3.9. Histogramas de estrato socioeconómico.

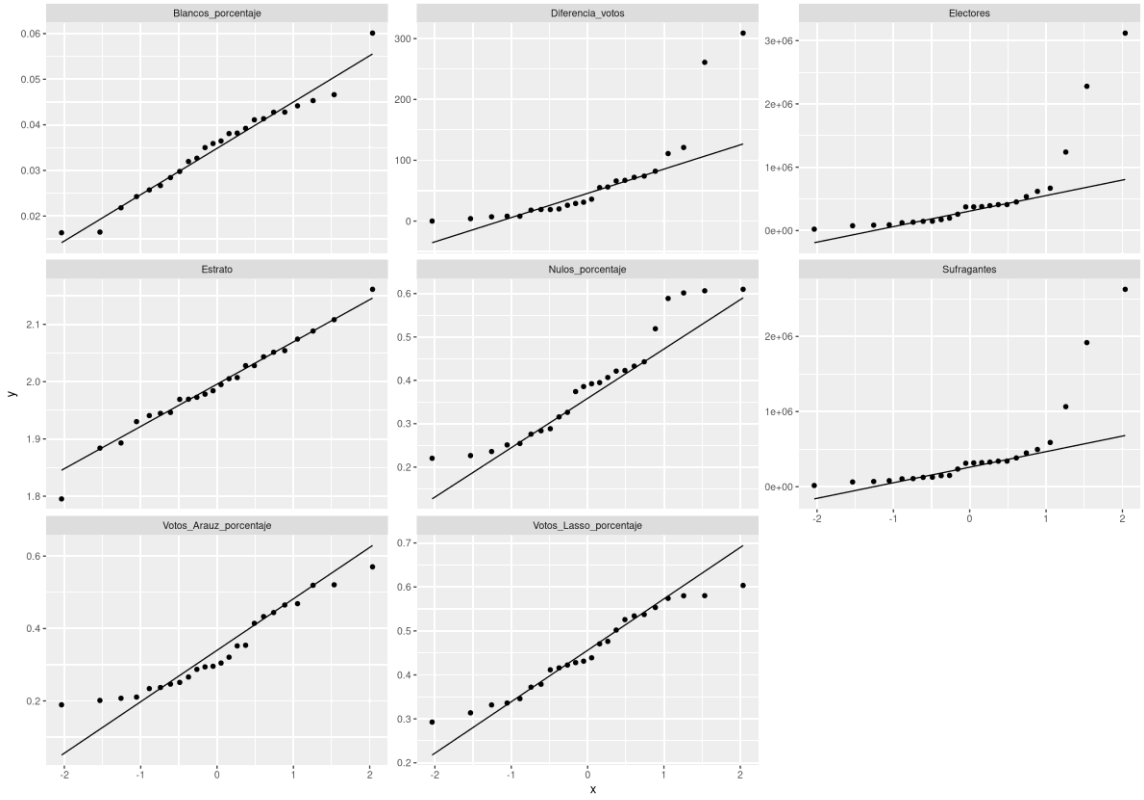

Figura 3.10. Gráficos QQ de votos, electores, sufragantes y estrato socioeconómico.

En la Figura 3.11 también se puede observar que los votos en blanco tuvieron correlaciones negativas con el número de sufragantes, electores y diferencia de votos; más votantes se correlacionaron con una menor cantidad de votos nulos y blancos (todo en porcentajes). Con el estrato socioeconómico la relación fue positiva, mientras más alto fue el estrato socioeconómico, más alto fue el número de votos en blanco. Los votos en blanco tuvieron una correlación negativa con los votos por Guillermo Lasso, más votos en blanco se relacionaron con menos votos para este candidato. Otra correlación interesante fue con Andrés Arauz, donde más votos en blanco se relacionaron con más votos para él.

Figura 3.11. Correlación entre variables, con estrato socioeconómico.

Adicionalmente, entre los votos de los candidatos, se tuvo una correlación negativa, si un candidato recibió más votos el otro recibió menos; pero su correlación (-0.82) no fue perfectamente negativa. El único candidato que tuvo correlación con el estrato socioeconómico es Andrés Arauz (-0.17); entre más alto fue el estrato socioeconómico, más bajo fue el porcentaje de votantes de Andrés Arauz. Esta correlación no fue replicada en el otro candidato, ya que no existió correlación entre el estrato socioeconómico y la cantidad de votos recibidos por Guillermo Lasso.

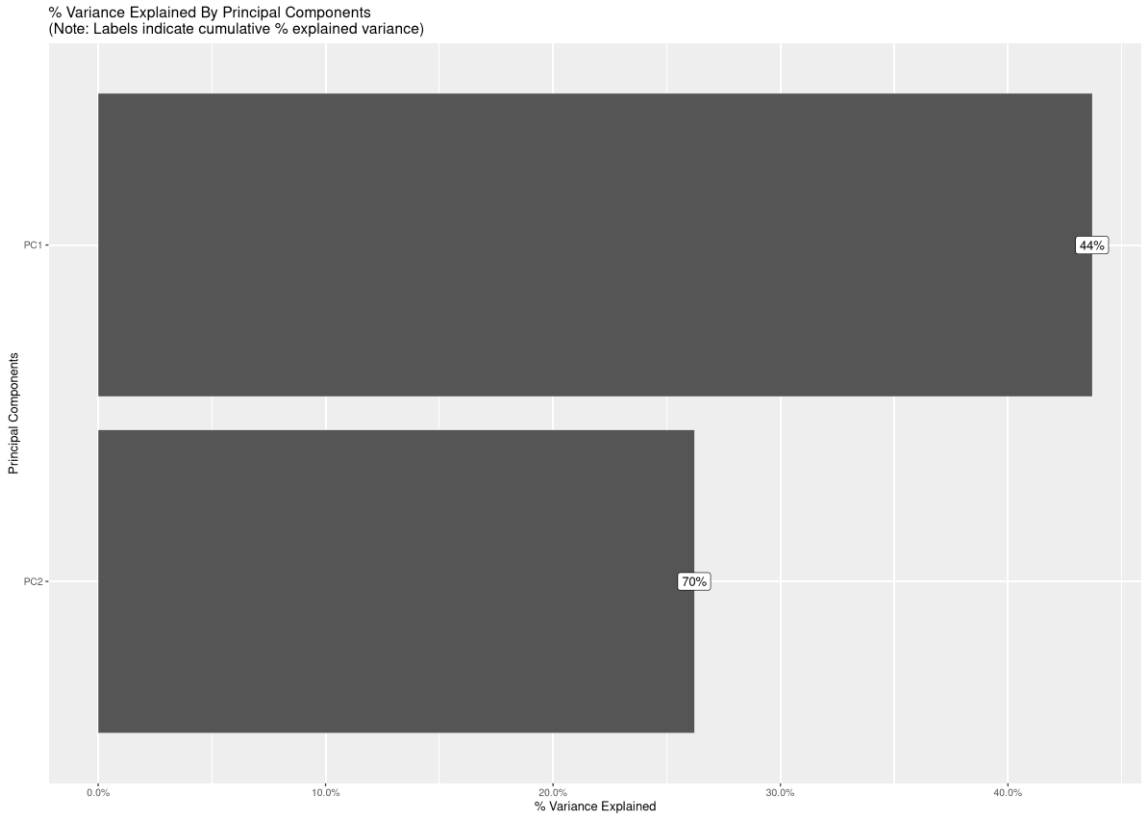

Figura 3.12. PCA de variables, incluyendo estrato socioeconómico.

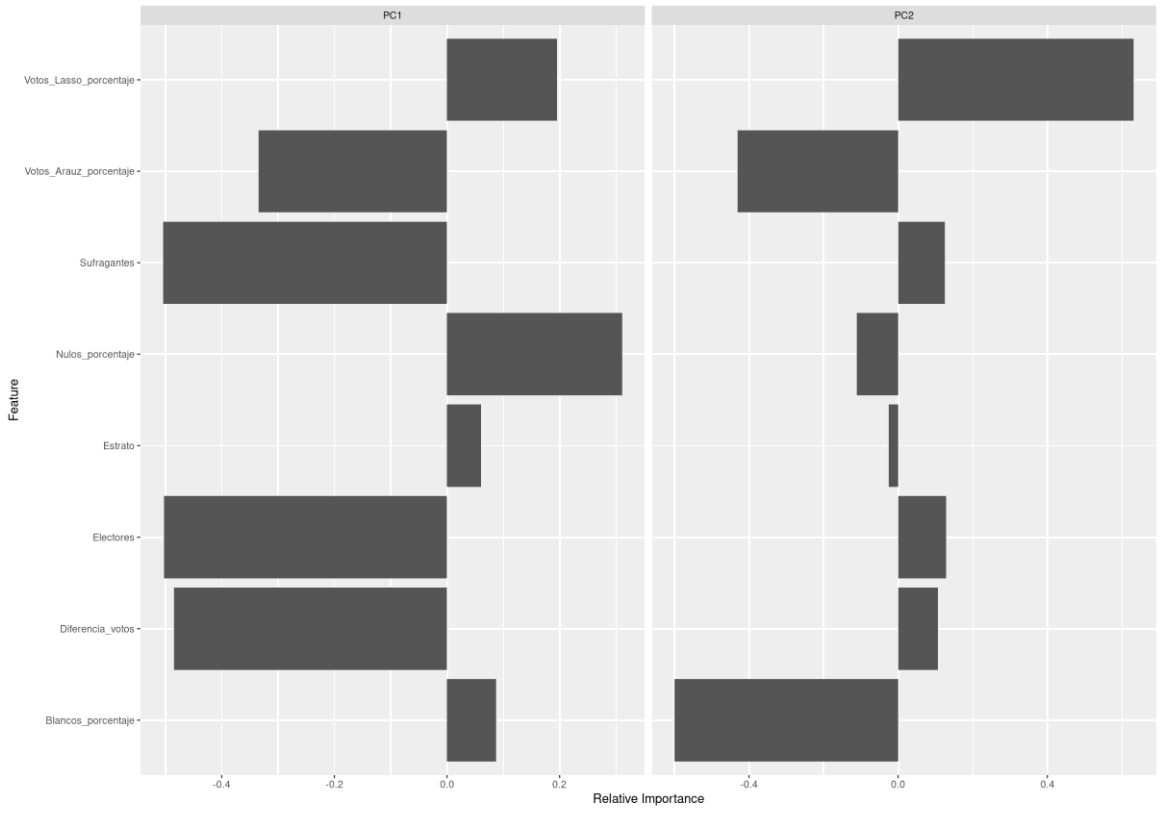

Figura 3.13. Importancia componentes, incluyendo estrato socioeconómico.

Finalmente, se puede ver en la Figura 3.12 que el análisis de componentes principales nos arrojó dos componentes, que representan el 70% de la varianza. Esto se observa de mejor manera en la Figura 3.13, donde está la importancia de cada variable en el cálculo. Se puede ver que cada variable representa componentes específicos, excepto en el caso de los votos de Arauz, que no fueron representados en los componentes.

Análisis computacional general

Sistema operativo

Se puede asumir, que las libertades de GNU/Linux fueron una de las principales razones para que los sistemas virtuales de Kaggle hayan sido diseñados usando este sistema operativo. Estos sistemas virtuales necesitan ejecutarse bajo entornos especializados, y con el uso de notebooks, por lo que es necesario poder personalizarlos constantemente y el uso de un software libre permite realizar esta personalización. Adicionalmente, en un entorno libre se puede desarrollar más fácilmente entornos de programación, por la facilidad en el manejo de los componentes necesarios para tener el comportamiento deseado. Además, al trabajar en máquinas virtuales se puede otorgar al usuario acceso root sin generar problemas, ya que el acceso se limita a esa máquina virtual. Tener un usuario root permite una gran cantidad de facilidades, como modificar, leer y mover archivos en diferentes sectores.

Infraestructura física

Tal y como se observó en la sección anterior, existen diferencias marcadas en memoria RAM y procesador entre el entorno GPU y el entorno CPU. Las diferencias en esta área vienen dadas por un ahorro en recursos de parte de Kaggle, dado que en un entorno de GPU los recursos de un procesador deberían ser menos necesarios. Otra razón para esta diferencia es el costo de un GPU, como se puede ver en el marco teórico el modelo de negocios de Kaggle se beneficia de los creadores de competencias. Por esta razón se incentiva un uso consciente de GPU en los demás usuarios, ya que un usuario que realice cálculos sin GPU, tendría un mejor desempeño en un entorno de CPU. Para el análisis realizado en este documento no se esperaría tener una diferencia muy grande en cuanto a los cálculos efectuados en un entorno u otro. Esto debido a que la mayoría de cálculos no se benefician del uso de un GPU. Adicionalmente, la cantidad de datos es baja, por lo que los cálculos se realizan de manera rápida sin necesidad del uso de un GPU.

Gracias al uso de Kaggle, se tiene acceso gratuito a esta infraestructura, de manera ilimitada para el caso de entornos de CPU, y con cuota semanal para entornos de GPU.

Gracias al modelo de negocios de Kaggle, que incentiva a los competidores a crear y probar modelos computacionales para entregar a sus clientes que crearon la competencia el mejor modelo posible. Por esta misma razón, Kaggle permite realizar cálculos largos y almacenamiento permanente de los resultados enviando una ejecución de manera asíncrona, con un límite de ejecución de 9 horas. Otra ventaja es que en Kaggle se tiene el entorno ya configurado previamente, para que funcione con las librerías y paquetes que ellos consideran que son los más utilizados en la ciencia de datos.

Posibilidades de computación paralela y distribuida

R tiene librerías listas para ser utilizadas, las cuales ocupan el GPU directamente en sus cálculos sin requerir configuración. Usando este tipo de librerías es sencillo realizar paralelización con el uso de un GPU en los entornos de Kaggle. Este tipo de computación es usada en Kaggle para competencias con redes neuronales y análisis de imágenes, ya que las tareas realizadas en esos casos son enormemente paralelizables por las operaciones entre matrices que son realizadas; además, utilizan una gran cantidad de datos.

Base de Datos

Los datos son cargados rápidamente hacia el entorno virtual en el que se trabaja gracias al uso de HDFS, el cual permite que los datos se carguen rápidamente desde los clústeres más cercanos, y de forma paralela. En el caso de los datos usados en este trabajo, la principal ventaja es la carga desde el clúster más cercano, ya que los ficheros tienen un tamaño pequeño (menos de 1GB). Adicionalmente, el hecho de que Kaggle permita utilizar cualquier formato para los datos, permite realizar el trabajo de manera más directa y muy similar a lo que se podría realizar en un entorno local.

Seguridad de los datos

En el caso de los datos del CNE, existe la posibilidad de obtener datos más específicos que indican la parroquia donde los votantes están registrados, lo cual sí podría llegar a ser un riesgo en determinados casos; pero los datos usados en este trabajo no acceden a esa información. Por otro lado, los datos usados de la encuesta del INEC son generales sobre empleo, desempleo y subempleo de las parroquias, por lo que no se tiene un riesgo a la privacidad de las personas. Además, tanto los conjuntos de datos, como el documento reproducible, son públicos; por lo que la seguridad de Kaggle no es una preocupación. En el caso de que los datos, o los notebooks, fueran privados, tendrían una protección para ser accedidos a través del usuario que los creó, con su usuario y contraseña.

Lenguaje R

Como R es un lenguaje interpretado, se puede hacer análisis de datos de manera sencilla al ejecutar las líneas de forma independiente e ir experimentando con diferentes opciones para obtener la perspectiva que se desea. La desventaja de usar este tipo lenguaje radica en su desempeño más lento que el de un lenguaje compilado, por lo que es compensado a través del uso de paquetes compilados, que otorgan un mejor desempeño.

Análisis computacional específico

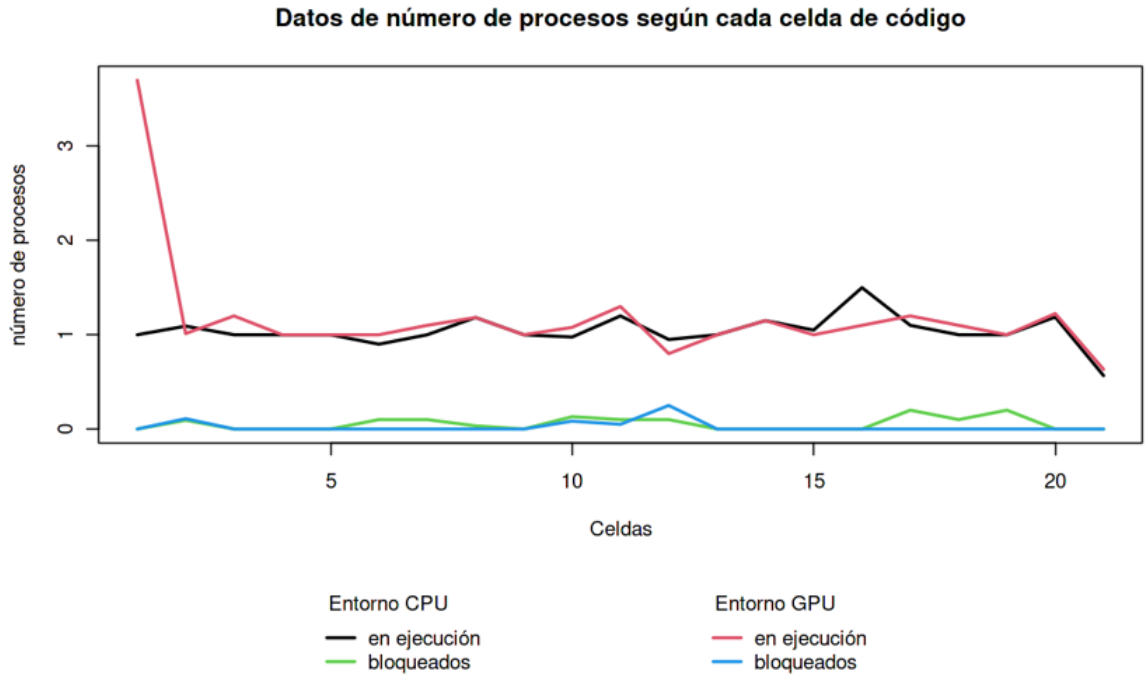

Figura 3.14. Número de procesos en cada celda de código.

Luego de obtener los datos para el análisis computacional específico y graficarlos se procedió a interpretarlos. Los gráficos se muestran con las celdas en el eje x, empezando desde la celda 1, y las métricas en el eje y. En la Figura 3.14 se observa el número de procesos del sistema en las celdas, donde lo primero que se puede notar es que la cantidad de procesos en ejecución al momento de arrancar el sistema fue alta para el entorno de GPU. Esto tal vez se haya debido a que los entornos de GPU requieren de una mayor configuración al momento de arrancar. Además, el número de procesos en ejecución siempre se mantuvo más alto que el número de procesos bloqueados, en ambos entornos. Destacan también la carga de datos (celdas 3 y 8), la impresión del reporte (celda 11), la generación del primer histograma (celda 16), y la creación del segundo reporte (celda 20). Adicionalmente, el número de procesos bloqueados destaca en la realización de clústeres (celda 12), en la creación de histogramas (celdas 17 y 18), y en la creación de dataframes

(celda 19). Éste fue el comportamiento esperado dado que se ejecutó un único proceso en la mayoría de operaciones.

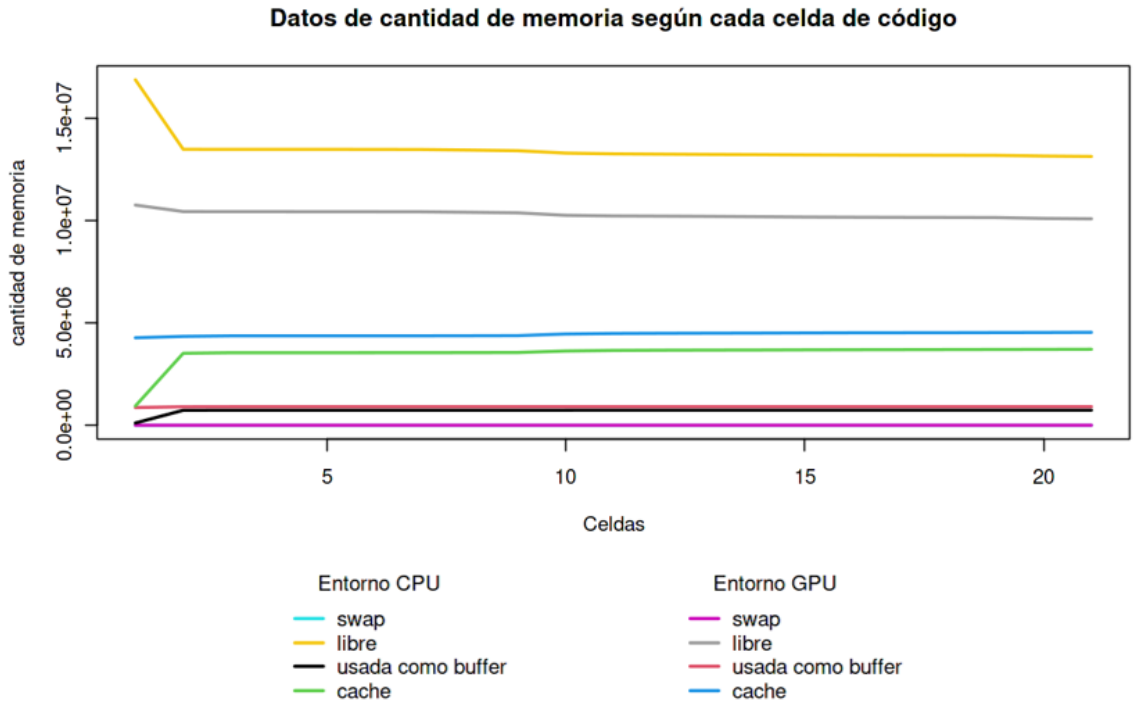

Figura 3.15. Memoria en cada celda de código.

Ciertos procesos requirieron esperar la terminación de otros antes de continuar la ejecución. Se observa esto al ver un movimiento combinado entre procesos bloqueados y ejecutándose. Por ejemplo, luego de la celda 16, donde la librería parece haber paralelizado algunas operaciones realizadas en el dataframe; en la celda 17 del entorno CPU la cantidad de procesos bloqueados aumentó. Por otro lado, en el entorno de GPU se observa un comportamiento similar en la celda 11, donde el número de procesos en ejecución aumentó, para luego disminuir drásticamente en la celda 12, y aumentar el número de procesos bloqueados. Otro dato obtenido es el de la cantidad de memoria, observado en la Figura 3.15, donde la memoria usada se mantuvo estable la mayor parte del tiempo en ambos entornos. La parte más destacable es cuando se realiza la carga de librerías (celda 2), donde la memoria libre se redujo a medida que se incrementó la memoria de cache y la usada como buffer. En el entorno de CPU esto fue más pronunciado, probablemente debido a que el entorno de GPU cargaría ciertas librerías directo a la memoria dedicada, y no se aprecia en este gráfico.

En el caso de swap, la Figura 3.16 muestra que se mantuvo en 0 durante todo el tiempo de ejecución. La cantidad de memoria usada fue muy pequeña en comparación con la

disponible, como se vio en la Figura 3.15. Debido a esto, no se requirió hacer swap de memoria en ningún momento de la ejecución.

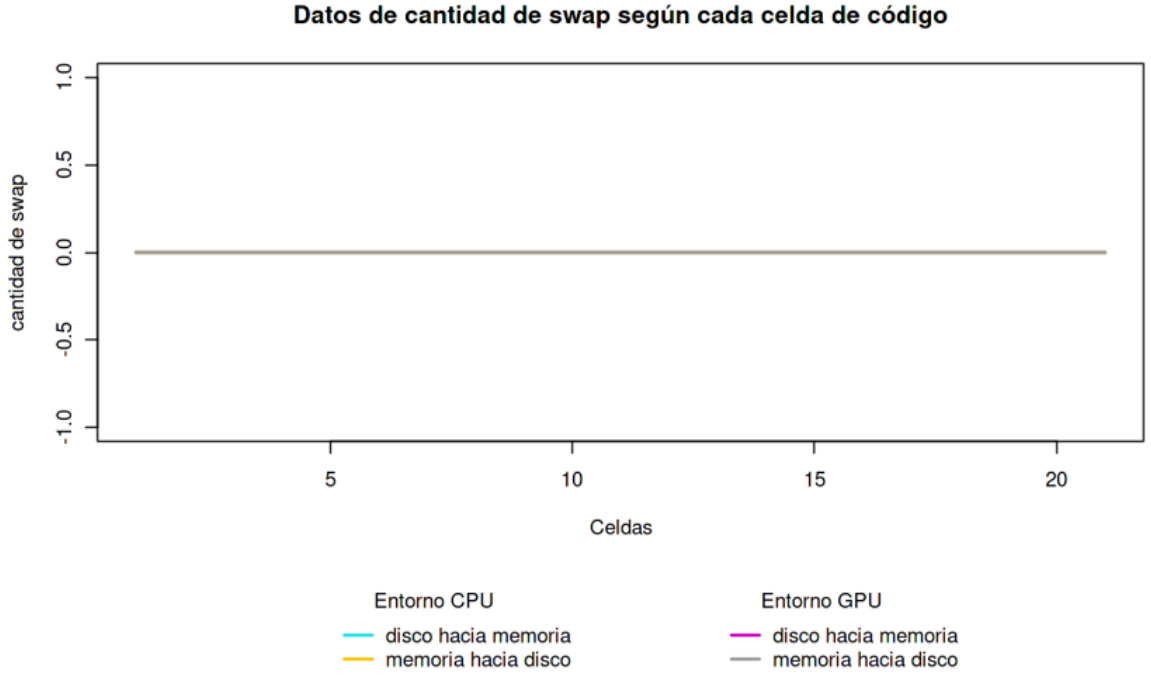

Figura 3.16. Swap en cada celda de código.

También se puede observar el número de bloques transferidos entre memoria y disco, según cada celda. En la Figura 3.17 se observa que, al arrancar el sistema en el entorno de GPU, se envió cerca de 10000 bloques a disco. En el análisis de correlaciones y de componentes principales (celda 10) se observa una gran cantidad de datos que fueron recibidos de disco. Sin embargo, en la celda 20, la cual tiene un análisis similar, ya no sucedió ese comportamiento. Esto es un indicativo de que la librería usada para realizar este análisis no había cargado todos sus componentes al iniciar, más bien los cargó cuando se requirió su ejecución. Por esta razón al ejecutar la celda 20 ya no se observó ese comportamiento, donde los componentes ya habían sido cargados y fueron usados directamente de memoria, haciendo uso del cache. Un bloque viene dado por 1024 bytes, por lo que los 10000 bloques apenas constituyen 10 MB; siendo que la Figura 3.15 posee escalas en GB, no se puede observar este comportamiento. También se observa que la cantidad de bloques enviados desde la celda 11 en adelante aumentó, con un pico notable para el entorno de GPU en la celda 17. El incremento constante para estas celdas se debió a que se realizó la creación de los gráficos del análisis exploratorio, los cuales fueron guardados en disco una vez terminados. En el caso de la celda 17 no es claro por qué hubo

una diferencia tan notable entre los entornos de GPU y CPU, pudo deberse a un gráfico en particular que el GPU haya creado usando mayor cantidad de datos.

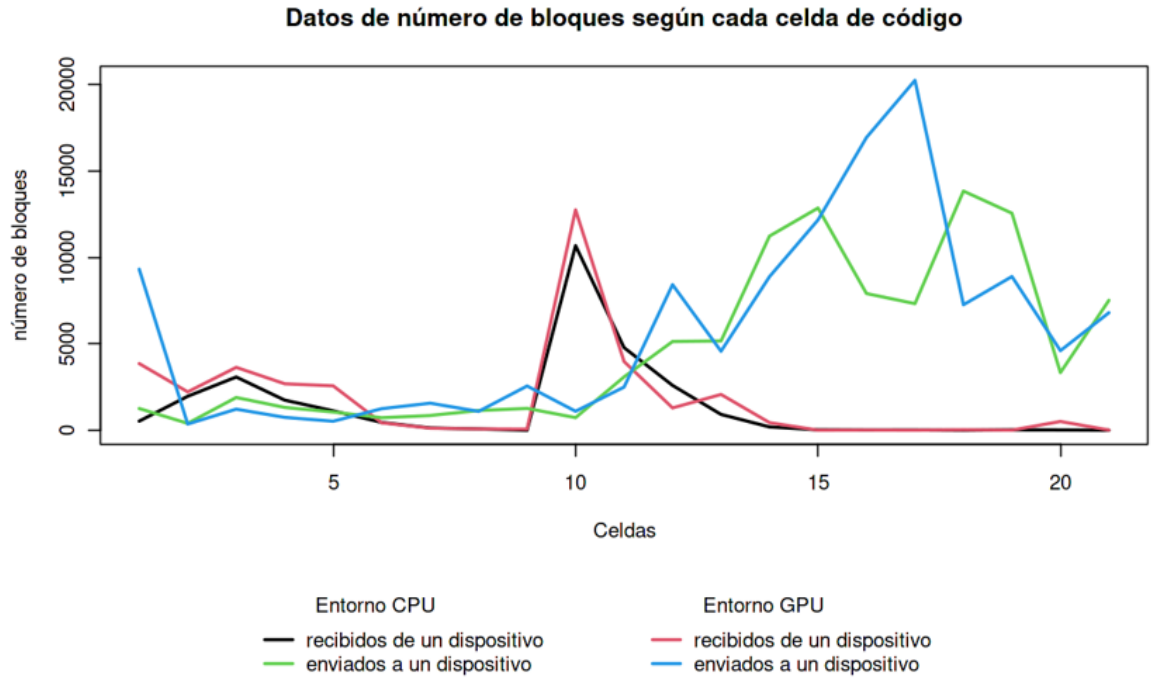

Figura 3.17. Bloques transferidos en cada celda de código.

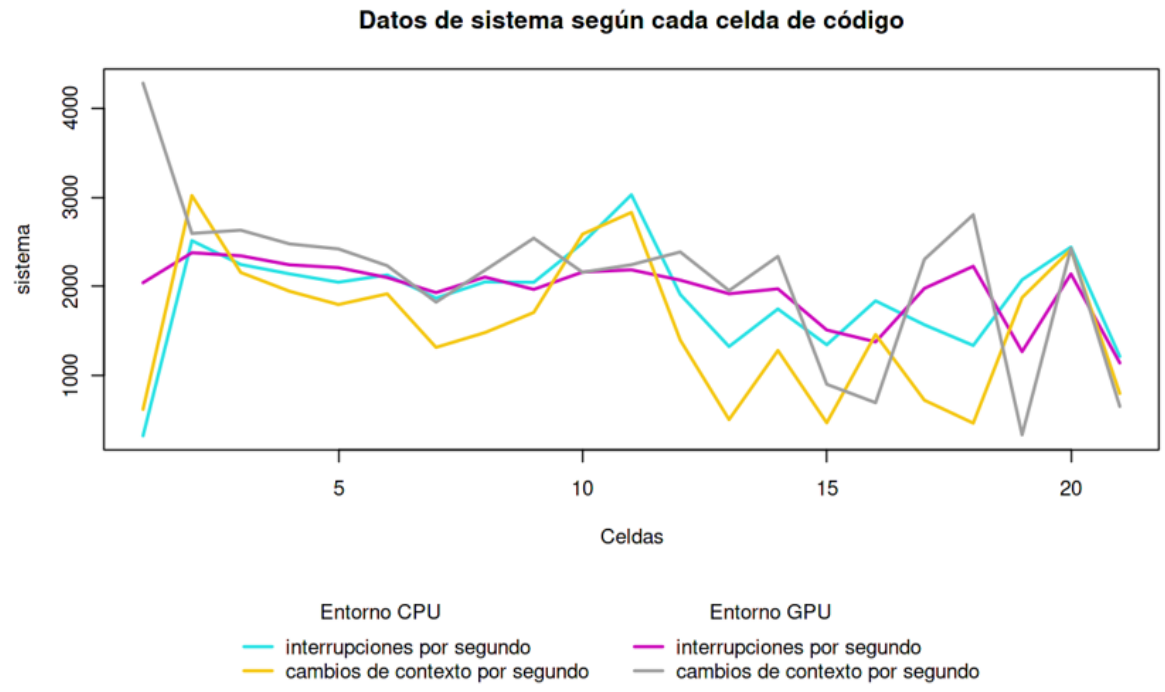

Figura 3.18. Datos de sistema en cada celda de código.

En la Figura 3.18 se tiene los datos del sistema, donde los cambios de contexto por segundo comenzaron con fuerza en el entorno de GPU. Nuevamente esto pudo ser debido a que ese entorno requiere de mayor configuración inicial. Hay varios puntos de interés en la Figura 3.18, como la impresión del reporte (celda 11), donde que destacó en el entorno de CPU; o las operaciones en dataframe (celda 9), donde destacó el entorno GPU. Es importante recordar que en los entornos de GPU hubo un procesador más lento y con menos núcleos, por lo que se esperaba más interrupciones y cambios de contexto. Se puede observar que en la creación de gráficos (celdas 16, 17 y 18) esto se volvió más pronunciado. Sin embargo, en los análisis de componentes principales y correlaciones (celdas 10 y 20), el entorno GPU tuvo menos cambios de contexto e interrupciones que el de CPU. Esto se debió a que estas operaciones sí estaban haciendo uso del GPU internamente, lo cual lo es verificable más claramente en la Figura 3.20.

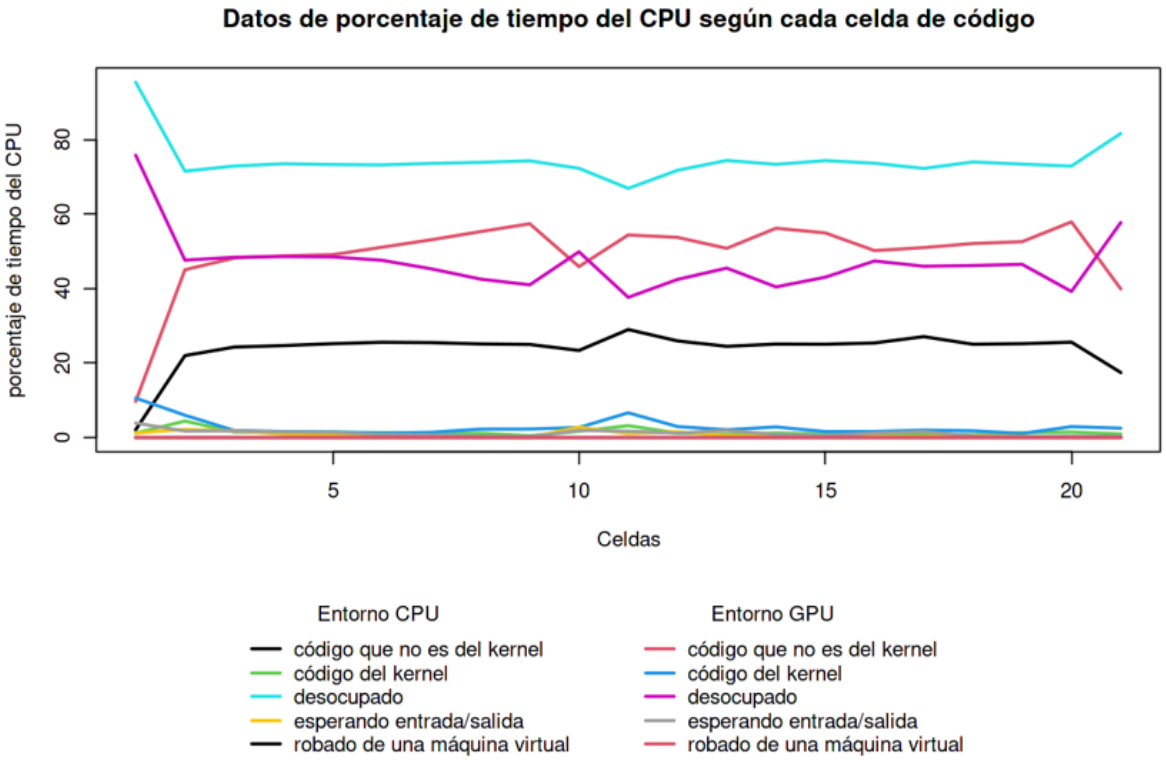

Figura 3.19. Porcentaje de tiempo del procesador en cada celda de código.

Para los datos de la Figura 3.19 se puede observar que en el entorno de CPU el procesador pasó la mayor parte del tiempo desocupado, con el resto del tiempo ocupado principalmente en código que no es de kernel. Para el entorno de GPU, se tuvo una repartición de alrededor de 50% en cada uno de estos porcentajes. Los demás porcentajes fueron muy pequeños para distinguirlos, a excepción de lo que se alcanza a ver al inicio, por las configuraciones iniciales, y en la celda 11 y 21 al momento de imprimir los reportes

de análisis. Se confirmó nuevamente que el entorno de CPU tenía un procesador más potente, el cual no fue aprovechado, ya que pasó desocupado la mayor parte del tiempo. Otras celdas que destacaron son la celda 9 del entorno GPU, y las celdas 11, 14 y 20 en ambos entornos. La celda 14 se encargó de imprimir datos del dataframe, después de calcular aquellos que tengan un mayor número de electores, por lo que el incremento en el código que no es de kernel pudo deberse a esta razón. Finalmente se tiene que la celda 20 tuvo un incremento en la ejecución de código que no es de kernel para el entorno de GPU, el cual vino de la ejecución del análisis del último dataframe creado en la celda 19.

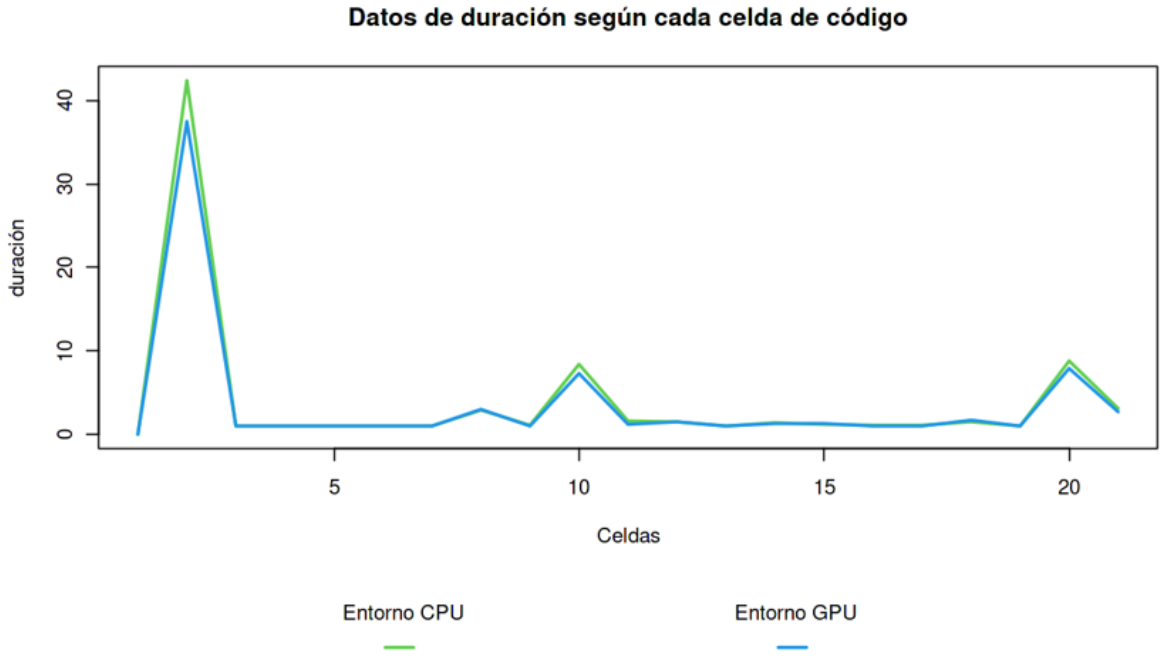

Figura 3.20. Duración en cada celda de código.

Finalmente, para la duración, se observa en la Figura 3.20 4 puntos donde la duración destacó; mientras que los demás puntos indicaron una duración de 1 segundo (que en la mayoría de casos sería de menos, por la precisión de las marcas de tiempo). El entorno GPU realizó las operaciones ligeramente más rápido en la carga de librerías y la generación de los reportes, teniendo la misma duración en las demás tareas. Así se pudo verificar que la librería DataExplorer utilizó el GPU en el entorno de forma automática, ya que las celdas donde fue ejecutada son las que redujeron su duración. Además, se puede observar que paralelizar las operaciones de análisis de componentes principales y comparación de cuantiles fue más efectivo que usar un procesador con mejor capacidad.

3.2. Conclusiones

Análisis exploratorio

I) Primera hipótesis:

- La diferencia entre el total de votos y el total de sufragantes se debió a errores efectuados por los miembros de las juntas receptoras de voto.

Como fue definido en la metodología, la primera hipótesis, de ser cierta, tendría consecuencias específicas, las cuales son analizadas a continuación.

- El número de sufragantes y la cantidad de votos que no coinciden tuvieron un comportamiento similar, de hecho, fue casi idéntico en su distribución. Si aumentaba el número de sufragantes en un determinado lugar, también aumentaba el número de votos que no coinciden. Lo cual es indicativo de que los errores fueron más frecuentes mientras más sufragantes existían. La correlación entre las variables fue muy alta (0.86), tal y como se pudo apreciar en la Figura 3.3.
- El porcentaje de votos que no coincidieron con respecto a la cantidad de sufragantes fue muy bajo (inferior a 1%), y fue similar en todas las figuras, tal y como se puede observar en las Figuras 3.6, 3.7 y 3.8.
- Las parroquias que se comportaron de forma anómala lo hicieron por razones ajenas a la cantidad de votos que no coincidieron (diferencia entre votos y sufragantes). En este caso la anomalía se dio por el número de electores, que además estaba correlacionada fuertemente con las demás variables (1 con el número de sufragantes y 0.86 con la diferencia de votos), y genera que la parroquia haya sido detectada como anómala. Esto pudo ser apreciado en las Tablas 3.1 y 3.2.

Debido a lo establecido anteriormente, se puede concluir que la hipótesis fue comprobada. En otras palabras, la diferencia entre el total de votos y el total de sufragantes sí se debió a errores efectuados por los miembros de las juntas receptoras de voto.

II) Segunda hipótesis:

- Los votantes tuvieron preferencia por algún candidato según el estrato socioeconómico al que pertenecían.

Como fue definido en la metodología, la segunda hipótesis, de ser cierta, tendría consecuencias específicas, las cuales son analizadas a continuación.

- No se pudo apreciar una correlación entre la cantidad de votos de cada candidato y el estrato socioeconómico medio de la provincia. Existió una correlación negativa de -0.17 entre los votos de Arauz y el estrato socioeconómico (si el estrato era bajo, había más votos para Arauz), pero no existió una correlación entre el estrato socioeconómico y los votos para Lasso (no importa el estrato al momento de ver la cantidad de votos para Lasso) ya que tuvo un valor de 0.01. Esto se puede observar en la Figura 3.11.

Por lo expresado anteriormente, se puede concluir que la segunda hipótesis queda parcialmente refutada. Es decir, los votantes tuvieron preferencia parcial por algún candidato en específico según el estrato socioeconómico al que pertenecían, si pertenecían a estrato bajo era probable que tuvieran preferencia por Andrés Arauz, pero la preferencia por Guillermo Lasso no estuvo ligada a ningún estrato. En ambos casos el análisis está incompleto, ya que los datos usados fueron de la segunda vuelta electoral, donde los candidatos recibieron también votos de personas que no los habían elegido como su primera opción.

Análisis computacional general

Kaggle es una excelente opción gratuita para desarrollar investigación y trabajos en inteligencia artificial, análisis de datos, aprendizaje de máquina y otras disciplinas relacionadas. Sin embargo, debemos tener en cuenta la limitación impuesta por el uso de un entorno virtual al cual no se tiene acceso completo, si no únicamente a través del notebook. Además, estas limitaciones también vienen debido al modelo de negocios enfocado a los creadores de competencia y no a los competidores o usuarios casuales, los cuales tienen acceso gratuito. Adicionalmente, el uso de un Sistema Operativo como el de Linux en este entorno nos da la libertad necesaria para realizar cálculos y análisis, pero el entorno no permite el acceso a un nivel muy profundo, lo cual impide el desempeño de funciones específicas, como instalar o mejorar herramientas del sistema operativo. Por otro lado, el uso de R en este entorno es facilitado enormemente por el hecho de que viene configurado, y podemos beneficiarnos de la facilidad que R otorga para analizar datos e instalar librerías.

Análisis computacional específico

Hipótesis

- Las celdas 3, 6, 8, 10, 20 y 12 consumieron más recursos computacionales que las demás al momento de ejecutarse el análisis exploratorio.

Consecuencias

- Los recursos computacionales consumidos mostrarán valores más altos en las celdas 3, 6, 8, 10, 20 y 12.

Según lo analizado, las celdas 3, 6 y 8 destacaron en las medidas de entrada y salida, observable en las Figura 3.15 y 3.17. Mientras que las celdas 10, 20 y 12 destacaron en procesos bloqueados (Figura 3.14), cantidad de bloques enviados (Figura 3.17), tiempo del CPU dedicado a código que no es de kernel (Figura 3.19), interrupciones y cambios de contexto por segundo (Figura 3.18). Sin embargo, hubo otras celdas que destacaron, como las celdas que realizaban creación de gráficos para mostrarlos (celdas 11 en adelante). También destacó la primera celda en los entornos de GPU, por la configuración inicial requerida, y la celda de instalación y carga de librerías. Por todo esto podemos concluir que nuestra hipótesis fue parcialmente comprobada. Los puntos tomados en cuenta en la hipótesis se cumplieron, pero hubo puntos que se encontraron al realizar el análisis que no fueron contemplados inicialmente, por lo que la hipótesis inicial estaba incompleta.

Cumplimiento de objetivos

Objetivo general

Se demostró correctamente los conocimientos y habilidades adquiridas en las diferentes asignaturas del currículo de la carrera. Esto se logró a través de la realización del presente proyecto de análisis exploratorio de datos, y los correspondientes análisis computacionales.

Objetivos específicos

1. Se pudo determinar el comportamiento de variables socioeconómicas y electorales de las elecciones generales del 2021. Adicionalmente, se logró definir hipótesis sobre las correlaciones encontradas entre dichas variables, las cuales fueron comprobadas en unos casos, y refutadas en otros. Todo esto se realizó en un documento reproducible.

2. Se pudo analizar la infraestructura y organización del entorno utilizado al momento de realizar el objetivo anterior. Esto se logró al analizar la configuración del sistema y comprender el funcionamiento de sus partes.

3. Se pudo analizar el rendimiento de la infraestructura utilizada para la realización del TIC con el uso de la herramienta vmstat, el cual se incluyó en un segundo documento reproducible.

3.3. Recomendaciones

Análisis exploratorio

Para realizar un análisis exploratorio de forma adecuada se debe mantener un enfoque claro y evitar posibles preconcepciones que terminen llevando al análisis en una dirección equivocada. De no hacerlo así se podría ignorar ciertos datos que darían una importante perspectiva a las hipótesis planteadas inicialmente. Esta recomendación surge del hecho que inicialmente se esperaba encontrar correlaciones más fuertes entre el estrato socioeconómico y los candidatos de cada provincia. Gracias al uso del método hipotético-deductivo, esta preconcepción fue mitigada al momento de realizar el análisis.

Como investigación futura se recomienda analizar a mayor profundidad las correlaciones encontradas entre otras variables a nivel de provincia, como el porcentaje de votos nulos y el porcentaje de votos por Arauz, o el porcentaje de votos blancos. Existen muchas explicaciones posibles para esa correlación y otras, pero quedan fuera del alcance de este trabajo. Adicionalmente, se recomienda realizar un análisis de los datos de la primera vuelta en las elecciones generales del 2021, ya que los datos obtenidos en cuanto a preferencia están incompletos, debido a que los candidatos recibieron también votos de personas que no los eligieron como primera opción.

Análisis computacional general

El entorno de Kaggle podría recibir algunas mejoras, como mayor memoria RAM; la cual podría llegar a ser útil para casos de procesamiento con más datos. Adicionalmente, se podría añadir otro tipo de paquetes para realizar el monitoreo de los recursos del sistema. El entorno virtual además se puede beneficiar de una interfaz más clara para realizar el despliegue de reportes, ya que los paquetes de R acostumbran a emitirlos en formato HTML.

Análisis computacional específico

Para la recolección de este tipo de datos se recomienda un sistema automatizado, debido a la gran cantidad de datos que se requiere para estabilizar las muestras. Para este trabajo se recolectó datos de 10 ejecuciones por cada entorno, pero una cantidad mucho mayor nos entregaría una mayor calidad de datos. Adicionalmente se recomienda buscar la manera de obtener una mayor precisión en las medidas, para muchas celdas de código la duración fue inferior a 1 segundo, lo cual dificultó el emparejamiento con los datos de la

herramienta vmstat. Finalmente, se recomienda utilizar directamente entornos para análisis de datos como es el caso de R, ya que permite ejecutar los distintos análisis de forma más rápida, directa y automatizable.

4. REFERENCIAS BIBLIOGRÁFICAS

[1] Consejo Nacional Electoral, «CNE,» 2022. [En línea]. Available: https://www.cne.gob.ec/estadisticas/bases-de-datos/. [Último acceso: 11 06 2022].

[2] INEC, «Banco de Datos Abiertos,» 2022. [En línea]. Available: https://aplicaciones3.ecuadorencifras.gob.ec/BIINEC-war/index.xhtml. [Último acceso: 15 06 2022].

[3] Kaggle, «Kaggle,» 2022. [En línea]. Available: https://www.kaggle.com/. [Último acceso: 15 06 2022].

[4] A. Volkens, J. Bara y I. Budge, «Data quality in content analysis: the case of the comparative manifestos project,» *Historical Social Research,* vol. 34, nº 1, pp. 234-251, 2009.

[5] M. Laver, «Analysing structures of party preference in electronic voting data.,» *Party Politics,* vol. 10, nº 5, pp. 521-541, 2004.

[6] L. H. Jacintho, T. P. da Silva, A. R. Parmezan y G. E. Batista, «Analyzing spatio-temporal voting patterns in Brazilian elections through a simple data science pipeline,» *Journal of Information and Data Management,* vol. 12, nº 1, pp. 31-47, 2021.

[7] P. Godfrey-Smith, «Glossary,» de *Theory and reality*, Chicago, The University of Chicago Press, 2003, p. 236.

[8] Gobierno de la república del Ecuador, «Historia de la Función Electoral,» CNE, [En línea]. Available: https://www.cne.gob.ec/historia-de-la-funcion-electoral/. [Último acceso: 13 08 2022].

[9] Gobierno de la república del Ecuador, «Reporte de resultados preliminares,» CNE, [En línea]. Available: https://resultados2v.cne.gob.ec/Resultados/VentanaReporte. [Último acceso: 13 08 2022].

[10] Gobierno de la república del Ecuador, «La Institución,» INEC, [En línea]. Available: https://www.ecuadorencifras.gob.ec/la-institucion/. [Último acceso: 13 08 2022].

[11] INEC, «Recálculo de las estadísticas de empleo y pobreza,» Quito, 2021.

[12] R. Rufilanchas, «On the origin of Karl Pearson's term "histogram",» *Revista Estadística Española,* vol. 192, pp. 29-35, 2010.

[13] I. S. Ufimtsev y T. J. Martinez, «Graphical Processing Units for Quantum Chemistry,» *Computing in Science and Engineering,* vol. 10, pp. 26-34, 2008.

[14] R. Colburn, Sams Teach Yourself CGI in 24 Hours, Sams Publishing, 2003.

[15] Microsoft, «Byte Ordering,» 2019. [En línea]. Available: https://docs.microsoft.com/en-us/openspecs/office_file_formats/ms-doc/dc135247-563f-42e3-8b60-1d0c36402840. [Último acceso: 13 08 2022].

[16] Free Software Foundation, Inc., «GNU Hurd,» 2019. [En línea]. Available: https://www.gnu.org/software/hurd/. [Último acceso: 13 08 2022].

[17] R. Stallman, «Linux y el sistema GNU,» 2022. [En línea]. Available: https://www.gnu.org/gnu/linux-and-gnu.es.html. [Último acceso: 13 08 2022].

[18] The Apache Software Foundation, «HDFS Architecture Guide,» 2008. [En línea]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. [Último acceso: 13 08 2022].

[19] Canonical Ltd., «Ubuntu Manpage: vmstat,» 2019. [En línea]. Available: https://manpages.ubuntu.com/manpages/bionic/es/man8/vmstat.8.html. [Último acceso: 20 08 2022].

[20] R. Ihaka, «A Brief History,» [En línea]. Available: https://cran.r-project.org/doc/html/interface98-paper/paper_2.html. [Último acceso: 20 08 2022].

[21] W. Cleveland, The Elements of Graphing Data, Hobart Press, 1994.

[22] P. Peres-Neto, D. Jackson y K. Somers, «How many principal components? stopping rules for determining the number of non-trivial axes revisited,» *Computational Statistics & Data Analysis,* vol. 49, nº 4, pp. 974-997, 2005.

[23] Wolfram Research, Inc., «Correlation Coefficient,» 2022. [En línea]. Available: https://mathworld.wolfram.com/CorrelationCoefficient.html. [Último acceso: 20 08 2022].

[24] J. B. MacQueen, «Some Methods for classification and Analysis of Multivariate Observations,» de *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1*, California, University of California Press, 1967, pp. 281-297.

5. ANEXOS

ANEXO I

Documento reproducible, con código, del análisis exploratorio ubicado en la página web de Kaggle.

https://www.kaggle.com/code/jdragonxherrera/tic-analisis

ANEXO II

Documento reproducible, con código, del análisis computacional específico ubicado en la página web de Kaggle.

https://www.kaggle.com/code/jdragonxherrera/tic-analisis-computacional