ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

ANÁLISIS ESTADÍSTICO DE LOS PROCESOS ELECTORALES EN ECUADOR

ANÁLISIS EXPLORATORIO ESTADÍSTICO DE LOS PROCESOS ELECTORALES EN ECUADOR

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

JOSELYN SELENA TACO PULUPA

joselyn.taco@epn.edu.ec

DIRECTOR: LUIS ENRIQUE MAFLA GALLEGOS

enrique.mafla@epn.edu.ec

DMQ, septiembre 2022

CERTIFICACIONES

Yo, JOSELYN SELENA TACO PULUPA 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.

JOSELYN SELENA TACO PULUPA

Certifico que el presente trabajo de integración curricular fue desarrollado por JOSELYN SELENA TACO PULUPA, bajo mi supervisión.

LUIS ENRIQUE MAFLA GALLEGOS
DIRECTOR

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.

JOSELYN SELENA TACO PULUPA

LUIS ENRIQUE MAFLA GALLEGOS

DEDICATORIA

Este trabajo se lo dedica a:

A mis padres, por su amor y paciencia, quienes han sacrificado todo por darme los medios necesarios para cumplir todos mis sueños, por su fortaleza y tenacidad, lo que me ha impulsado a luchar y no dejarme vencer hasta conseguir lo que quiero. Estoy agradecida con ellos por inculcar en mí los mejores valores, por ser mi ejemplo para seguir y enseñarme a no temer a las adversidades que se presenten en mi camino, porque Dios está conmigo.

A mi hermana por su cariño, paciencia y apoyo en cada momento, en cada paso que he dado. Finalmente, a mis abuelos quienes fueron mis segundos padres, gracias por su amor, sus consejos, oraciones y palabras de aliento, gracias por hacer de mí una mujer de bien, alguien con principios y valores. Espero, donde quiera que estén se sientan orgullosos de mí.

AGRADECIMIENTO

Agradezco a Dios por tolas bendiciones en mi vida, por guiarme a lo largo de mi existencia, ser el apoyo y fortaleza en aquellos momentos de dificultad y de debilidad.

Gracias a mis padres Susana y Carlos por ser los principales promotores de mis sueños, por confiar y creer en mis expectativas, por los consejos, valores y principios que me han inculcado.

Agradezco a mis docentes de la Escuela de Politécnica Nacional, por haber compartido sus conocimientos a lo largo de mi preparación como profesional. De manera especial, al ingeniero Enrique Mafla Gallegos tutor de este trabajo de integración curricular quien me ha guiado con paciencia, y rectitud como docente durante todo el desarrollo del presente trabajo.

.

ÍNDICE DE CONTENIDO

RESUMEN

El Trabajo de Integración Curricular (TIC) tiene como objetivo poner en práctica los conocimientos adquiridos durante la carrera, al realizar un análisis exploratorio de datos relacionado a los procesos electorales y a las condiciones socioeconómicas del Ecuador. Los datos procesados son obtenidos en los sitios Web del CNE e INEC. Además, se integra en este trabajo un análisis computacional integral de los sistemas computacionales utilizados, así como una evaluación de rendimiento a dichos sistemas. Para todo lo mencionado anteriormente, el trabajo sigue una metodología de investigación mixta (cualitativa y cuantitativa) para cada uno de los análisis. El TIC utiliza el lenguaje estadístico R para realizar el análisis exploratorio de datos. Al finalizar con el trabajo los resultados muestran la presencia de una relación estadísticamente significativa entre los resultados electorales y el nivel de analfabetismo en Ecuador; así como, la evidencia de un consumo alto de procesamiento y almacenamiento al realizar el análisis exploratorio de datos.

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

ABSTRACT

The Curricular Integration Work (TIC) puts into practice the knowledge acquired during the career, by performing exploratory analysis of data related to the electoral processes and the socioeconomic conditions of Ecuador. The processed data are obtained from the CNE and INEC websites. In addition, a comprehensive computational analysis of the computational systems used is integrated with this work, as well as a performance evaluation of these systems. For all of the above, the work follows a mixed research methodology (qualitative and quantitative) for each analysis. The TIC uses the statistical language R to perform the exploratory data analysis. At the end of the work, the results show a statistically significant relationship between electoral results and the level of illiteracy in Ecuador; as well as the evidence of high consumption of processing and storage when performing the exploratory data analysis.

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

1 DESCRIPCIÓN DEL COMPONENTE DESARROLLADO

En el trabajo de integración curricular (TIC) realizamos un análisis exploratorio de datos relacionado a los procesos electorales y a las condiciones socioeconómicas del Ecuador. El TIC incluyo un análisis computacional integral de los sistemas computaciones de procesamiento, almacenamiento y entrada-salida, utilizados para la realización del mencionado análisis exploratorio y una evaluación del rendimiento de dichos sistemas computacionales. El análisis computacional y la evaluación de rendimiento cumplieron con el objetivo de integración curricular del TIC.

El trabajo de integración curricular estuvo compuesto por 3 partes: análisis exploratorio de datos, análisis integral de los sistemas computacionales y evaluación de rendimiento. En el análisis exploratorio de datos examinamos las relaciones estadísticas entre los resultados electorales a nivel provincial y las variables que representan la situación económica de la población. Para llevar a cabo dicho análisis, utilizamos los datos disponibles en los sitios Web del Consejo Nacional Electoral (CNE) y del Instituto Nacional de Estadística y Censos (INEC). Para desarrollar el análisis exploratorio de datos utilizamos el lenguaje de programación estadística R. Esta parte del TIC, la realizamos utilizando las técnicas de investigación reproducible [1]. El Anexo II contiene el enlace del documento reproducible correspondiente

En la segunda parte, realizamos un análisis computacional integral de los sistemas computacionales que utilizamos para el procesamiento, almacenamiento y entrada – salida durante el TIC. Realizamos dicho análisis a nivel de arquitectura de computadores, sistema operativo, middleware y ambiente de programación. Además, investigamos otras plataformas, a más de R, utilizadas para análisis de datos.

En la tercera parte evaluamos el rendimiento de una infraestructura local y una infraestructura en la nube, para ejecutar las tareas del análisis exploratorio. La infraestructura local es un computador personal. Para la implementación de la infraestructura en la nube utilizamos los servicios de computación en la nube de AWS. En esta evaluación medimos el rendimiento de R para la ejecución de las tareas computacionales del análisis exploratorio de datos

En el TIC obtuvimos los siguientes resultados, en sus tres componentes.

El análisis exploratorio de datos evidenció que existe una relación estadísticamente significativa entre el nivel de analfabetismo y los resultados electorales. Es decir, mientras

el total de votos válidos aumenta el porcentaje de personas analfabetas se reduce significativamente.

El análisis computacional integral evidenció que para la obtención de datos necesitamos ampliar nuestro ancho de banda en el caso de que el TIC requiera utilizar grandes volúmenes de datos. Además, para la tarea de limpieza y depuración de datos necesitamos una memoria RAM con al menos 16 GB de capacidad; así como utilizar un sistema de interconexión que evite el paso por memoria RAM, es decir un acceso directo a cache.

Finalmente, la evaluación de rendimiento encontró que el consumo de recursos computacionales en la infraestructura local fue del 71.65%. A diferencia del consumo en una infraestructura en la nube que fue del 30%. Además, se observó que la infraestructura local llevo a cabo todas las tareas del análisis exploratorio con éxito. Pero, necesito mayor capacidad de recursos computacionales que una infraestructura en la nube. La cual, como se mencionó anteriormente, no necesito utilizar grandes recursos computacionales.

El documento está organizado de acuerdo con la estructura del TIC. Es decir, cada uno de los capítulos y secciones del documento tiene tres partes; una para cada una de las componentes del TIC. En lo que resta de la Introducción presentamos los objetivos generales y específicos, el alcance y el marco teórico del TIC. En el capítulo dos, describimos la metodología que empleamos para ejecutar el TIC. En el capítulo tres, analizamos los resultados obtenidos y presentamos nuestras conclusiones y recomendaciones.

1.1 Objetivo general

Demostrar los conocimientos, habilidades, valores y aptitudes adquiridas en las diferentes asignaturas del currículum de la carrera, mediante la realización de un proyecto de análisis exploratorio de datos sociales, económicos y electorales de Ecuador.

1.2 Objetivos específicos

1. Encontrar relaciones estadísticas entre variables electorales y socioeconómicas del Ecuador mediante un análisis exploratorio de datos.
2. Determinar de manera general los requerimientos de procesamiento, almacenamiento y entrada-salida, de los procesos computacionales utilizados para llevar a cabo el mencionado análisis.

3. Evaluar el rendimiento de una infraestructura local y en la nube para análisis exploratorio de datos.

1.3 Alcance

Tomando en cuenta las limitaciones de tiempo, recursos computacionales, de datos y de información definí el siguiente alcance para el TIC:

Análisis exploratorio: Durante el análisis exploratorio de datos, definí hipótesis sobre tendencias y correlaciones de los datos, y realicé pruebas estadísticas preliminares de dichas hipótesis. Dicho análisis incluyo los procedimientos de adquisición, depuración y estructuración de datos. Para realizar el análisis exploratorio utilice técnicas de investigación reproducible [1]. El TIC no incluye el análisis definitivo de dichas hipótesis para su confirmación o rechazo.

Análisis computacional integral: En el TIC realicé el análisis computacional integral de los requerimientos de procesamiento, almacenamiento y entrada – salida a nivel de arquitectura, sistema operativo, middleware y ambiente de programación, para realizar análisis estadísticos de datos. Dicho análisis integral estuvo basado en el curriculum de la ACM [2] y en la malla curricular de la carrera de ICC de la EPN [3].

Evaluación de rendimiento: En el TIC evalué el rendimiento de una infraestructura local y una infraestructura en la nube para la ejecución de las tareas del análisis exploratorio de datos. La infraestructura en la nube estuvo basada en los servicios gratuitos de computación en la nube de AWS.

1.4 Marco teórico

En esta sección, presentamos los marcos teóricos de las tres componentes del TIC: análisis exploratorio de datos, análisis computacional integral y evaluación de rendimiento.

1.4.1 Análisis exploratorio de datos

Distribución de frecuencia de datos

Se denomina frecuencia al número de casos u ocurrencias que contiene cada categoría o variable en una base de datos. La distribución de frecuencia enlista esta información del número de ocurrencias por cada categoría. A partir de la información enlistada se puede hallar patrones de ocurrencia en los datos [4].

Dispersión estadística de datos

La dispersión estadística es el valor en que una distribución de datos se aleja o acerca a la media aritmética. Se ocupan las medidas de dispersión para observar la variabilidad o dispersión de los datos. Las medidas de dispersión son: rango (diferencia entre el menor y mayor medida de la distribución), desviación media (promedio de las desviaciones de cada dato con respecto a la media), coeficiente de variación o de variación de Pearson (valor de la dispersión relativa de los datos), varianza y desviación típica [5]

Correlación de variables de Pearson

Una correlación de variables mide la relación lineal entre dos variables cuantitativas. Se puede crear una tabla que contenga los coeficientes de relación entre variables. El coeficiente estará en un rango entre 1 y –1. Si el coeficiente es mayor a 0, indica que la correlación es positiva, es decir que las variables tienden a incrementarse juntas. En cambio, si es menor que 0, la correlación es negativa, lo que significa que los valores de una variable tienden a incrementarse mientras que los valores de la otra variable decrecen. Finalmente, si el coeficiente es igual a 0, se puede decir que no hay relación lineal entre las variables [6].

Investigación reproducible

La reproducibilidad está relacionada al concepto de replicación, el cual es el acto de repetir un método científico y poder obtenerlos resultados similares. Entonces, es un ambiente reproducible si los datos y el código de la investigación se encuentran disponibles para cualquier interesado. De tal manera que puedan ser capaces de obtener los mismos resultados de la investigación. Además, para que la investigación sea reproducible, cada paso debe estar detallado con claridad y estar debidamente documentada de principio a fin [1].

Revisión sistemática de literatura

En el TIC realizamos una revisión sistemática de literatura. En los resultados de dicha revisión encontramos estudios directamente relacionados a nuestro trabajo. A continuación, presentamos un resumen de dichas investigaciones

La investigación *Economía, política social y Twitter: análisis de las emociones negativas en cuatro elecciones presidenciales latinoamericanas a través del LIWC* [7], utiliza un software que analiza las emociones de 25 959 tweets de cuatro campañas presidenciales de Argentina, Perú, Ecuador y Chile entre los años 2015 y 2017. Tras finalizar la investigación encontraron para la comunidad Andina la existencia de varias emociones

como: el 2.64% emociones negativas en contra de Moreno y 2.52% en contra de Lasso. Un 0.99 % de enfado en contra de Moreno y un 0.92% en contra de Lasso.

La investigación *Ecuador TV como medio de propaganda en las elecciones presidenciales de la era Correa (2007-2017)* [8], utiliza un análisis cuantitativo y cualitativo para concluir la presencia de sesgos de propaganda política a favor del candidato oficialista y en contra de los rivales. Al concluir la investigación encontraron que, de un total de 95 piezas de propaganda analizadas, el 71,5% de ellas presenta propaganda a favor del candidato Correa. El 28.4% restante de las piezas presentan un contenido neutral.

1.4.2 Análisis computacional integral

Arquitectura y organización de computadores

La arquitectura de un computador es el conjunto de elementos tanto de hardware como de software que conforman el computador e influyen de manera directamente en las funciones y diseño de un computador. Entre los elementos asociados están: el procesador (encargado de la gestión y control de las operaciones del computador), memoria (se encarga del almacenamiento de información de los programas y datos necesarios para su ejecución), entrada – salida (se encarga de la comunicación y la transferencia de datos entre el computador y los dispositivos externos) [9].

Sistemas operativos

Un sistema operativo es el principal programa que se ejecuta en todo computador. Es el conjunto de programas que comparten el mismo propósito de administrar y extender las capacidades de los sistemas de información. Las capacidades son: el procesamiento realizado por la CPU, el almacenamiento de información que lleva a cabo la memoria y los dispositivos de entrada – salida [10].

Computación concurrente, distribuida y paralela

La computación concurrente ejecuta múltiples tareas de forma interactiva. Es decir, no realizan una tarea a la vez de forma secuencial. Diferentes procesos o equipos intercambiar y acceden a un mismo recurso en periodos de tiempo separados [11].

La computación paralela ejecuta múltiples instrucciones de manera simultánea. Existen varios niveles de paralelismo como: a nivel de bit, instrucción, datos y tareas [12].

La computación distribuida permite que equipos independientes trabajen en conjunto con el mismo propósito. Los recursos de estos sistemas son concurrentes por lo tanto se utilizan técnicas de exclusión para el acceso [13].

Lenguaje de programación

Un entorno de desarrollo integrado (IDE) es un sistema software que combina herramientas de desarrollo en una interfaz de usuario grafica (GUI). Un IDE cuenta con un editor de código fuente, herramientas de interpretación, compilación y depuración de código. Además, estos pueden estar dedicados a uno o varios lenguajes de programación [14].

Las estructuras de datos son una manera en que se organizan los datos en una computadora. Esto con el fin de permitir realizar operaciones y manipulaciones a un conjunto de datos. Entre las estructuras te datos conocidas están: vectores (almacena datos homogéneos y de tamaño fijo), arrays (permite almacenar datos homogéneos y estructurados), matrices (permite almacenar datos en 2 dimensiones) y data frames (admite datos heterogéneos) [15].

1.4.3 Evaluación de rendimiento

Gestión de capacidad

Tiene como propósito que los servicios y recursos de una infraestructura de TI logren el desempeño estimado, satisfaciendo la demanda actual y futura de recursos. Dentro de esta práctica incluye los siguientes subprocesos: gestión de capacidad del negocio, gestión de capacidad del servicio y gestión de capacidad del componente [16].

Planificación de la capacidad

Es una de las actividades de la gestión de capacidad. Es la actividad responsable de predecir y pronosticar las capacidades y necesidades de recursos de una infraestructura de TI. Todo esto mediante la recopilación de capacidades del entorno por medio del uso de herramientas de medición [16].

2 METODOLOGÍA

En este capítulo describimos las metodologías utilizadas para la ejecución de las tres componentes del proyecto, incluyendo el enfoque y tipo de investigación.

La primera parte del TIC es un análisis exploratorio de datos con un enfoque cuantitativo y de tipo exploratorio – experimental. Se procesan los datos obtenidos en los sitios Web del CNE e INEC. Al final del proceso de limpieza y depuración de datos se establecerán hipótesis de relación entre las diferentes variables sociales, económicas y electorales del Ecuador.

En la segunda parte del TIC es un análisis integral de los requerimientos computacionales de procesamiento, almacenamiento y entrada – salida a nivel de arquitectura, sistema operativo, middleware y ambiente de programación, con un enfoque cualitativo y de tipo descriptivo – explicativo.

Por último, la evaluación de rendimiento la realizamos con un enfoque cuantitativo y de tipo exploratorio – experimental. Evaluamos el rendimiento computacional de las tareas realizadas durante el análisis exploratorio de datos, de las dos infraestructuras local y distribuida. La evaluación de la infraestructura local la realizamos en un computador personal; mientras que la evolución de la infraestructura distribuida la realizamos en AWS.

2.1 Análisis Exploratorio

Para ejecutar esta parte del TI, seguimos el procedimiento descrito en [17]. Dicho procedimiento consta de los siguientes pasos

2.1.1 Formulación de la pregunta de investigación (hipótesis)

En el presente trabajo planteamos la siguiente pregunta de investigación:

- ¿El nivel de analfabetismo en el Ecuador influye en el tipo de voto realizado?

2.1.2 Obtención de datos

Los datos requeridos para contestar la pregunta de investigación fueron:

- Registro de los resultados electorales durante las elecciones de las dignidades de presidente y vicepresidente en Ecuador (segunda vuelta, 2021). Estos datos

deben tener la información del tipo de voto (blanco, nulo, valido), los votos deben estar separados por provincia, además del número de sufragantes registrados.

- Registro del número de personas alfabetos y analfabetos registrados en Ecuador (censo poblacional, 2010). Los datos deben estar separados por provincia

2.1.3 Limpieza y depuración de datos

Una vez extraídos los datos procedimos a realizar una limpieza y depuración de datos. Es decir, primero detectamos los registros corruptos o imprecisos de las bases de datos. Para luego proceder a su corrección o en el caso de ser necesario su eliminación.

A lo largo del análisis encontramos las siguientes inconsistencias en las bases de datos:

- Inconsistencia en las actas de la base de datos del CNE
- Registros en blanco en la base de datos del INEC

Limpieza y depuración de la base de datos CNE

Realizamos una búsqueda de registros con valor nulo en toda la base de datos. Después analizamos las variables y su relevancia para responder nuestra hipótesis. Por lo mencionado anteriormente, las variables que fueron eliminadas son: dignidad nombre (nombre de la dignidad a escoger, en este caso presidente y vicepresidente), provincia nombre, circunscripción código y circunscripción nombre, cantón nombre, parroquia código y parroquia nombre.

Entonces, las variables con mayor relevancia para seguir con nuestro análisis fueron:

- **Provincia código**: a cada provincia se le asigno un código (anexo 1)
- **Cantón código:** a cada cantón se le asigno un código
- **Código del candidato:** a cada candidato se le asigno un código. Para poder definir el código de cada candidato se utilizó el comando subset (), que nos permitió sumar los votos válidos y según los resultados pudimos identificar el código de cada candidato.
    - 1021: Lasso
    - 1030: Arauz
- **Sexo de la junta:** existen dos tipos de juntas receptoras del voto, una masculina y otra femenina

- **Número de sufragantes:** el número total de sufragantes
- **Votos Nulos:** total de votos nulos
- **Votos Blancos:** total de votos blancos
- **Votos Válidos:** total de votos validos

Después de definir las variables con las que continuaríamos el análisis, realizamos un conteo de votos. Esto con el fin analizar si el total de sufragantes es igual al total de votos registrados (total de votos válidos, blancos y nulos). Entonces primero, calculamos el total de los votos registrados, es decir sumamos lo votos blancos, nulos y votos válidos de cada candidato. Después, calculamos el total de sufragantes registrados por cada candidato. Una vez realizados dichos cálculos, pudimos identificar que el total de sufragantes es de 10829823 y el total de votos es de 10828723.

En consecuencia, pudimos decir que existe una variación entre el total de sufragantes y los votos registrados. Dicha diferencia la representamos en una columna, para poder identificar el origen de los votos faltantes. Entonces, para realizar lo antes mencionado separamos la base de datos en dos según el código del candidato, esto debido a que se observó que se encuentra registrado el mismo valor de votos nulos y blancos por candidato. Las bases separadas tomaron el nombre del código de cada candidato, es decir:

- X1021: datos del candidato Lasso
- X1030: datos del candidato Arauz

Para unir los datos utilizamos el comando merge(), que nos permite unir dos dataframes por columnas o filas en comun. Despues de realizar el proceso antes mencionado, podemos observar la nueba base de datos en la figura 1.

```
> head(dataFull)
  PROVINCIA_CODIGO CANTON_CODIGO JUNTA_SEXO SUFRAGANTES BLANCOS NULOS
1                1           260          F        1018      31   460
2                1           260          F       11047     114  2506
3                1           260          F       13607     171  3268
4                1           260          F        1501      35   570
5                1           260          F        1614      27   675
6                1           260          F        1633      48   774
  X1021.VOTOS X1030.VOTOS
1         252         275
2        5543        2883
3        6798        3369
4         260         636
5         442         472
6         437         374
```

Figura 1. Nuevo conjunto de datos separado por candidato

Despues, continuando con el proceso de limpieza y depuracion, realizamos las siguientes operaciones con los datos:

- Sumamos las columnas de los votos válidos de cada candidato y agregamos como nueva columna el resultado
- Eliminamos las columnas de los votos de cada candidato después del calculo
- Sumamos las columnas de votos blancos, nulos y votos validos y agregamos como nueva columna el resultado
- Calculamos la diferencia de votos entre la columna Sufragantes y Total de votos

Una vez realizado todo el anterior proceso, pudimos observar que existen anomalías en el conteo de votos. Los valores mayores a 1, son votos faltantes al número de sufragantes registrados, por otro lado, los valores negativos son votos excedentes. Todo lo mencionado lo podemos observar en la figura 2.

```
> head(dataFull)
  PROVINCIA_CODIGO CANTON_CODIGO JUNTA_SEXO SUFRAGANTES BLANCOS NULOS VALIDOS
1                1           260          F        1018      31   460     527
2                1           260          F       11047     114  2506    8426
3                1           260          F       13607     171  3268   10167
4                1           260          F        1501      35   570     896
5                1           260          F        1614      27   675     914
6                1           260          F        1633      48   774     811
  VOTOS_TOTALES DIF_VOTOS
1          1018         0
2         11046         1
3         13606         1
4          1501         0
5          1616        -2
6          1633         0
```

Figura 2. Diferencia de votos con respecto al total de sufragantes.

Limpieza y depuración de la base de datos INEC

En la base de datos proporcionada por parte del INEC, pudimos observar que existen registros con valores nulos. Por lo cual, procedimos a completar los registros con valor nulo, ya que solo se trata de un mal tipeo de datos.

Entonces, realizamos el siguiente proceso:

- Completamos los registros que poseen valor nulo por el valor que corresponda para las columnas de provincia y cantón
- Remplazamos los registros que contengan la palabra total por nulo, ya que solo se trata de un registro que contiene la suma total de cada provincia o cantón de ser el caso

Una vez que terminamos el proceso de completar los registros, el resultado de como quedo la base de datos lo podemos observar en la figura 3.

```
> baseINEC
# A tibble: 9,414 x 7
   Provincia Canton               Parroquia  Etnia  Alfabeto Analfabeto Total
   <chr>     <chr>                <chr>      <chr>  <chr>    <chr>      <chr>
 1 Azuay     CAMILO PONCE ENRIQUEZ CAMILO PO~ INDÍG~ 96       7          103
 2 Azuay     CAMILO PONCE ENRIQUEZ NA         AFROE~ 721      63         784
 3 Azuay     CAMILO PONCE ENRIQUEZ NA         MONTU~ 255      24         279
 4 Azuay     CAMILO PONCE ENRIQUEZ NA         MESTI~ 9173     582        9755
 5 Azuay     CAMILO PONCE ENRIQUEZ NA         BLANC~ 780      48         828
 6 Azuay     CAMILO PONCE ENRIQUEZ NA         OTRO/A 42       -          42
 7 Azuay     CAMILO PONCE ENRIQUEZ NA         Total  11067    724        11791
 8 Azuay     CAMILO PONCE ENRIQUEZ EL CARMEN~ NA     ALFABETO ANALFABETO Total
 9 Azuay     CAMILO PONCE ENRIQUEZ NA         INDÍG~ 6        2          8
10 Azuay     CAMILO PONCE ENRIQUEZ NA         AFROE~ 61       3          64
# ... with 9,404 more rows
```

Figura 3. Complicación de registros en base de datos INEC

Como se pudo observar en la figura 3, aún existen valores nulos que eliminar, además de corregir valores en otras columnas, entonces realizamos el siguiente proceso:

- Eliminamos la columna Parroquia
- Eliminamos de los registros con valor nulo que quedaron
- Remplazamos el signo “-” por “0”.
- Transformamos la columna Parroquia y Etnia por códigos lo cuales se encuentran en el Anexo 1 y Anexo 2.

Después verificamos si la suma de la columna total coincide con la suma de las columnas Alfabeto y Analfabeto respectivamente; y pudimos comprobar que no existe ninguna diferencia de valores.

2.1.5 Homogenización de datos.

Al finalizar con la sección de limpieza y depuración, comenzamos con el proceso de homogenización de datos. Es decir, vamos ha agrupar los datos por provincia en una sola tabla. Para ello seguimos los siguientes pasos:

- Agrupamos los datos por provincia en las bases CNE e INEC
- Creamos una nueva tabla con los datos agrupados
- Eliminamos columnas de cada base de datos.
- Eliminamos registros que no pertenezcan a las 24 provincias.
- Unimos las dos bases de datos.

Al finalizar con este proceso obtuvimos los siguientes datos, los cuales podemos observar en la figura 4.

Provincia	VBlancos	Nulos	Validos	Alfabetos	Analfabetos
1	10231	150921	333689	925460	66062
2	3126	43951	99039	210266	34000
3	4493	45328	99653	265986	36870
4	1841	24243	97554	215192	14188
5	7624	96296	223139	542374	84692
6	6093	88084	245239	471370	74346

Figura 4. Nuevo conjunto de datos homogenizados

2.1.6 Análisis de datos

Después de todo el proceso que realizamos en las anteriores secciones. Procedimos a generar histogramas para ver la distribución de los datos. De igual manera, generamos graficas entre dos variables para observar su relación.

Histogramas

En la figura 5, en el histograma de la variable DIF_VOTOS pudimos observar la existencia de votos negativos. Los cuales son votos excedentes registrados. Además, observamos una gran frecuencia de botos nulos y blancos.

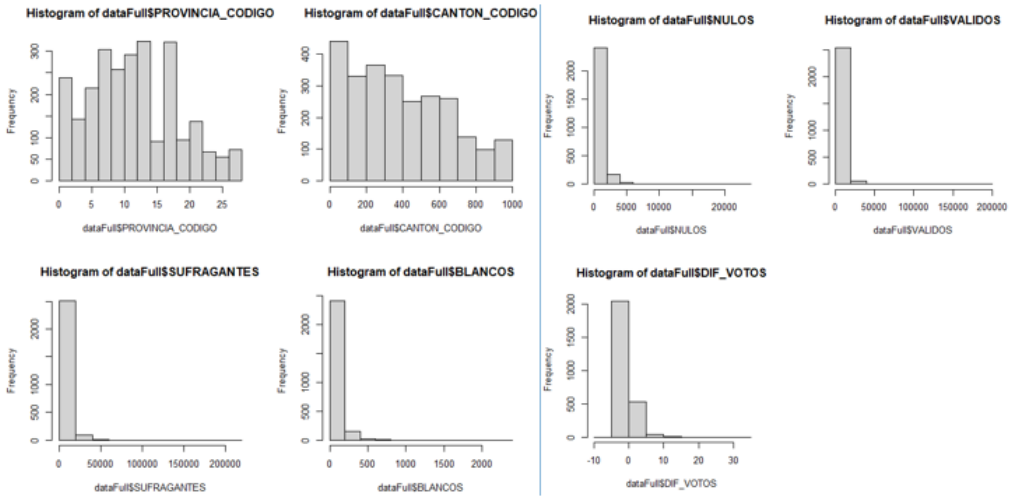

Figura 5. Histogramas de variables de votaciones electorales

En la figura 6, pudimos observar que el conteo de datos de registros de personas analfabetas es menor al de personas alfabetos. De igual manera, observamos que la frecuencia de datos es variada en la variable provincia.

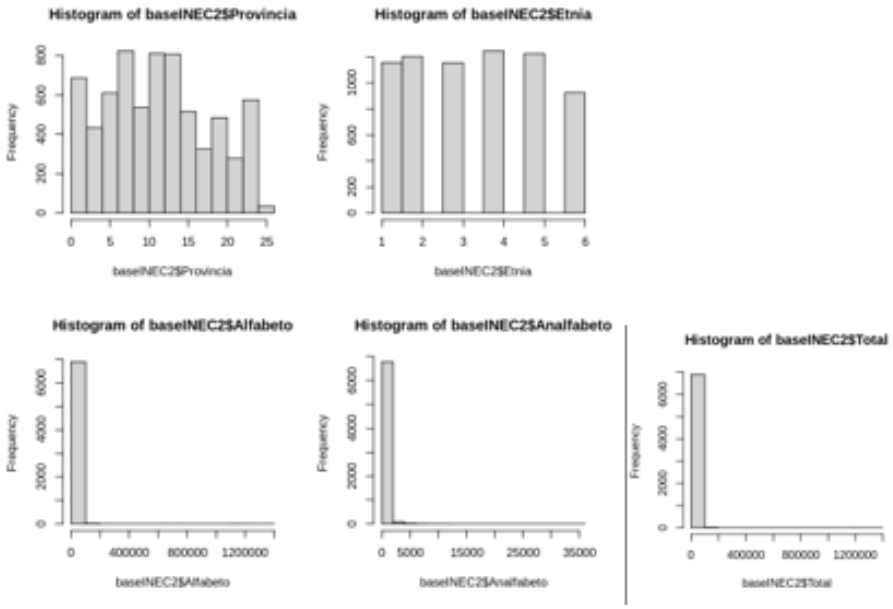

Figura 6. Histogramas de variables socioeconomicas

Matriz de Correlación

Las matrices de correlación nos permiten saber la relación lineal entre variables y para ello utilizamos el comando cor(), el cual nos permite genera dicha matriz. Por otro lado, para general la gráfica de la matriz de correlación, utilizamos el comando corrplot().

La figura 7, nos muestra la matriz de correlación de las variables de votaciones electorales. En la cual pudimos evidenciar una relación estadística entre los votos totales y votos blancos. De igual manera, existe una relación estadística entre los votos blancos y los votos validos

	PROVINCIA_CODIGO	CANTON_CODIGO	SUFRAGANTES	BLANCOS	NULOS
PROVINCIA_CODIGO	1.00	0.06	0.00	-0.04	-0.05
CANTON_CODIGO	0.06	1.00	-0.06	-0.03	-0.13
SUFRAGANTES	0.00	-0.06	1.00	0.92	0.95
BLANCOS	-0.04	-0.03	0.92	1.00	0.93
NULOS	-0.05	-0.13	0.95	0.93	1.00
VALIDOS	0.01	-0.05	1.00	0.91	0.93
VOTOS_TOTALES	0.00	-0.06	1.00	0.92	0.95
DIF_VOTOS	-0.02	-0.08	0.70	0.62	0.66

	VALIDOS	VOTOS_TOTALES	DIF_VOTOS
PROVINCIA_CODIGO	0.01	0.00	-0.02
CANTON_CODIGO	-0.05	-0.06	-0.08
SUFRAGANTES	1.00	1.00	0.70
BLANCOS	0.91	0.92	0.62
NULOS	0.93	0.95	0.66
VALIDOS	1.00	1.00	0.70
VOTOS_TOTALES	1.00	1.00	0.70
DIF_VOTOS	0.70	0.70	1.00

Figura 7. Matriz de correlación de variables de votaciones electorales

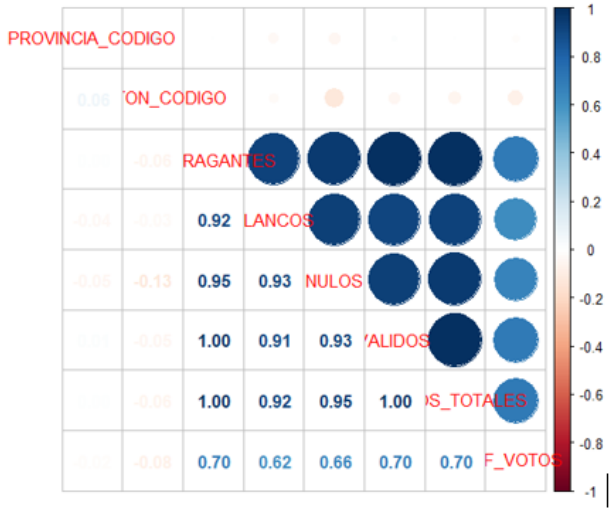

Figura 8. Grafica matriz de correlación de variables de votaciones electorales

La figura 9, nos muestra la matriz de correlación de las variables socioeconomicas. En la cual pudimos observar que existe relación estadística entre las variables analfabeto y alfabeto.

	Provincia	Etnia	Alfabeto	Analfabeto	Total
Provincia	1.00	0.00	0.01	-0.01	0.01
Etnia	0.00	1.00	0.01	-0.03	0.01
Alfabeto	0.01	0.01	1.00	0.87	1.00
Analfabeto	-0.01	-0.03	0.87	1.00	0.87
Total	0.01	0.01	1.00	0.87	1.00

Figura 9. Matriz de correlación de variables socioeconomicas

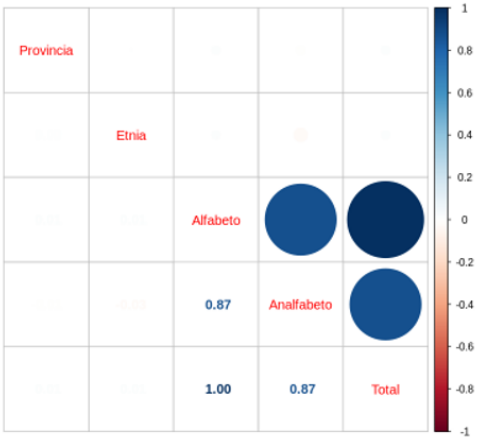

Figura 10. Grafica de correlación de variables socioeconomicas

Después de analizar las bases de datos por separado procedimos a analizar los datos homogenizados. Entonces, generamos la matriz de correlación (figura 11) y pudimos observar la existencia de una relación estadística entre los diferentes votos y las

personas analfabetas. Así como una relación estadística entre los votos válidos y los votos blancos.

	Provincia	VBlancos	Nulos	Validos	Alfabetos	Analfabetos
Provincia	1.00	-0.17	-0.20	-0.07	0.00	-0.14
VBlancos	-0.17	1.00	0.91	0.96	-0.12	-0.11
Nulos	-0.20	0.91	1.00	0.93	-0.15	-0.14
Validos	-0.07	0.96	0.93	1.00	-0.16	-0.19
Alfabetos	0.00	-0.12	-0.15	-0.16	1.00	0.88
Analfabetos	-0.14	-0.11	-0.14	-0.19	0.88	1.00

Figura 11. Matriz de correlación del conjunto de datos homogenizados

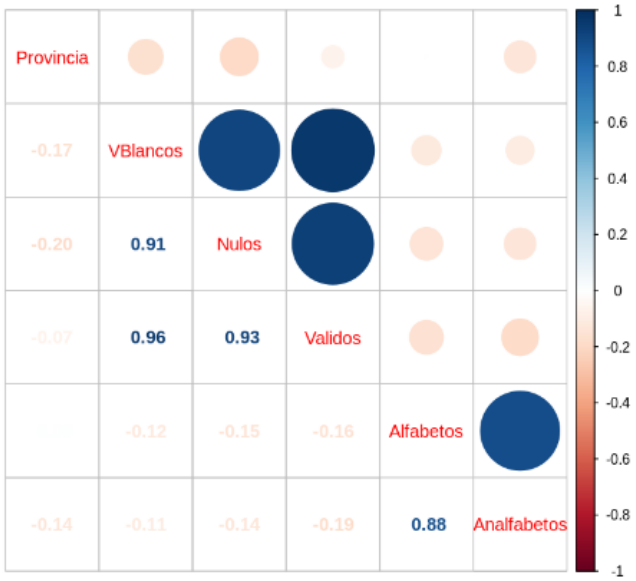

Figura 12. Grafica de correlación del conjunto de datos homogenizados

Al final de todo el proceso del análisis exploratorio de datos los resultados encontrados los presentamos en la sección de resultados.

2.2 Análisis Computacional Integral

En la segunda parte del trabajo de integración curricular, realizamos un análisis computacional integral de los sistemas computacionales utilizados para el procesamiento, almacenamiento y entrada – salida a nivel de arquitectura de computadores, sistema operativo y ambiente de programación. Esto con el fin de poner en práctica los conocimientos adquiridos en las asignaturas de la malla curricular de la carrera de ICC. A continuación, definimos las tareas que analizamos:

- Obtención de datos desde la web
- Limpieza y depuración de las bases de datos del CNE y del INEC

- Homogenización de datos
- Análisis de datos (histogramas, matrices de correlación y representación gráfica de las matrices de correlación)

Luego para cada tarea definimos los requerimientos computacionales generales. Con ello, analizamos el procesamiento, almacenamiento y entrada – salida de las tareas a nivel de arquitectura, sistema operativo y ambiente de programación como mencionamos anteriormente.

2.2.1 Arquitectura

En esta parte, definimos los requerimientos computacionales para cada tarea del análisis exploratorio que fue menciona anteriormente.

- Obtención de datos desde la web: En esta tarea los datos llegaron por la tarjeta de red y fueron al almacenamiento externo. Entonces los requerimientos generales fueron: ancho de banda, entrada – salida a almacenamiento en memoria, CPU y en entrada – salida de red.
- Limpieza y depuración de las bases de datos del CNE y del INEC: En esta tarea se cargan los datos del disco a la RAM. Entonces los requerimientos generales fueron: entrada – salida en disco, CPU y almacenamiento en RAM.
- Homogenización de datos: En esta tarea los datos ya están cargados en RAM. Entonces los requerimientos generales fueron: entrada – salida a RAM, CPU y almacenamiento en RAM.
- Análisis de datos: En esta tarea los datos ya están cargados en RAM. Entonces los requerimientos generales fueron: entrada – salida a RAM, CPU y uso de tarjeta gráfica.

Después analizamos los requerimientos de las tareas a nivel de arquitectura.

Procesamiento: Existen dos arquitecturas disponibles que nos sirven para realizar las tareas del análisis exploratorio. La arquitectura x86 (CISC) nos permite ejecutar instrucciones más complejas por lo que se necesita varios ciclos de reloj y por ello posee un mayor consumo de energía. Por otro lado, ARM (RISC) nos permite ejecutar instrucciones menos complejas (más cortas) por lo que suelen ser ejecutadas en un solo ciclo de reloj y por ello posee un menor consumo de energía.

Almacenamiento: Existen dos tipos de almacenamiento externo que tenemos a nuestro alcance. Los HDD poseen un alto consumo de energía y su velocidad de arranque es mayor a los SSD. Sin embargo, puede ser afectada por el magnetismo y eliminar los datos que guardemos. Por otro lado, un SSD tiene un consumo bajo de energía, además la velocidad de arranque que nos ofrece es menor a un HDD, pero su vida útil es reduce.

Para el almacenamiento interno tenemos: Un SDR SDRAM que nos ofrece ejecutar una instrucción de lectura – escritura por cada ciclo de reloj. Por otro lado, un DDR SDRAM ofrece el doble de velocidad y nos puede ayudar a realizar dos instrucciones de lectura – escritura por cada ciclo de reloj, además su número de pines paso de 168 a 184.

Entrada – Salida: El sistema de interconexión de datos tenemos dos tipos de transferencia disponibles: SATA posee un ancho de banda de 6Gbps y una velocidad de lectura – escritura de 600 Mbps. Por otro lado, PCIe posee un ancho de banda de 32 Gbps y una velocidad de lectura – escritura de 1Gbps.

2.2.2 Sistema Operativo

En esta parte, analizamos los requerimientos antes mencionados a nivel de sistema operativo.

Procesamiento: Podemos procesar las tareas de las siguientes maneras: concurrentemente podemos ejecutar las tareas de manera que todas progresen. La otra manera es paralelamente y las tareas se ejecutan de forma simultánea. También podemos tener en cuenta que, la ejecución por procesos consume más llamadas al sistema y por otro lado una sola llamada al sistema puede crear más de un hilo.

Almacenamiento: Podemos almacenar los datos de la siguiente manera: por archivo basándonos en una jerarquía de árbol podemos guardar los datos de manera organizada. Por bloque dividiremos los datos en grupos y se le asignara un identificador único a cada uno.

Entrada – Salida: Entre los sistemas de archivos existentes están: NTFS puede trabajar con archivos extensos, pero es inadecuado para particiones de menos de 400 Mb. Además, permite el cifrado de datos. HFS optimiza el espacio de almacenamiento, puede trabajar con un volumen máximo de 8 exbibytes y permite el cifrado de datos.

2.2.3 Ambiente de programación

En esta parte, analizamos los requerimientos antes mencionados a nivel de ambiente de programación.

Procesamiento: Las bibliotecas estadísticas y graficas de R son cargadas en memoria para su uso. Además, R funciona como un entorno temporal de trabajo.

Almacenamiento: Las estructuras de datos disponibles que podemos usar son: las matrices que nos permiten almacena datos estructurados del mismo tipo. Los data frames también nos permiten almacenar datos estructurados, con la diferencia de que las columnas pueden ser de distintos tipos de dato.

Entrada – Salida: Para mostrar los resultados obtenidos R utiliza los controladores gráficos que posee cada computadora. Así, como varias de sus bibliotecas.

2.3 Evaluación de rendimiento

En esta tercera parte del trabajo de integración curricular, realizamos una evaluación de rendimiento entre una infraestructura local y una infraestructura en la nube. Para este análisis utilizamos los resultados que se hallaron del análisis computacional integral de las tareas del análisis exploratorio.

Utilizamos las variables que son intensivas en el procesamiento, almacenamiento y entrada – salidas, encontradas en el análisis computacional integral. Entonces, primero comenzamos definiendo las infraestructuras que vamos a analizar.

Los detalles de las infraestructuras que utilizamos son:

Infraestructura local

La infraestructura local que utilizamos para el proceso de análisis es la siguiente:

Computador con las siguientes características:

- CPU 2.70GHz (4 CPUs) 2.90 GHz, 64 bits.
- RAM: 16,0 GB
- Sistema Operativo: Windows 10, Sistema operativo de 64 bits.

Después de que definimos la estructura local, utilizamos el **monitor de rendimiento** de Windows para medir la carga computacional al realizar las tareas del análisis exploratorio de datos.

Infraestructura en la nube

La infraestructura en la nube que utilizamos es una infraestructura de AWS con las siguientes características:

- Sistema Operativo: Ubuntu 18.04 LTS
- Amazon Elastic Compute Cloud (Amazon EC2)
- Amazon Elastic Block Store (EBS): 30 GB
- RStudio 1.3.1073

Para la infraestructura en la nube, utilizamos el monitoreo de Amazon Elastic Compute (EC2) y Cloud Watch para medir la carga computacional.

Para utilizar el monitoreo de EC2, primero activamos esta función en la instancia. Damos click derecho sobre la instancia, seleccionamos la opción de monitoreo y seleccionamos administrar monitoreo detallado. Después añadimos el monitoreo al panel de cloud watch.

Al finalizar todo el proceso de la evaluación de rendimiento los resultados encontrados los presentamos en la sección de resultados.

3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Resultados

En esta sección del documento se presentan los hallazgos encontrados. Estos están basados en las 3 partes de desarrollo del análisis como se mencionó en el alcance del documento. En la primera parte se detallan los resultados encontrados en el análisis exploratorio de datos. En la segunda parte están los resultados del análisis computacional general. Por último, se detallan los resultados del análisis computacional detallado

3.1.1 Resultados análisis exploratorio

En esta sección presentamos los resultados que obtuvimos al finalizar el análisis exploratorio de datos. Para ellos generamos gráficas, las cuales nos permiten resaltar aspectos de la distribución de los datos entre una o más variables cuantitativas.

Como primera grafica realizamos entre las provincias y votos nulos. En la figura 13 podemos observar que existe un mayor número de votos nulos en la provincia con código 9 (Guayas)

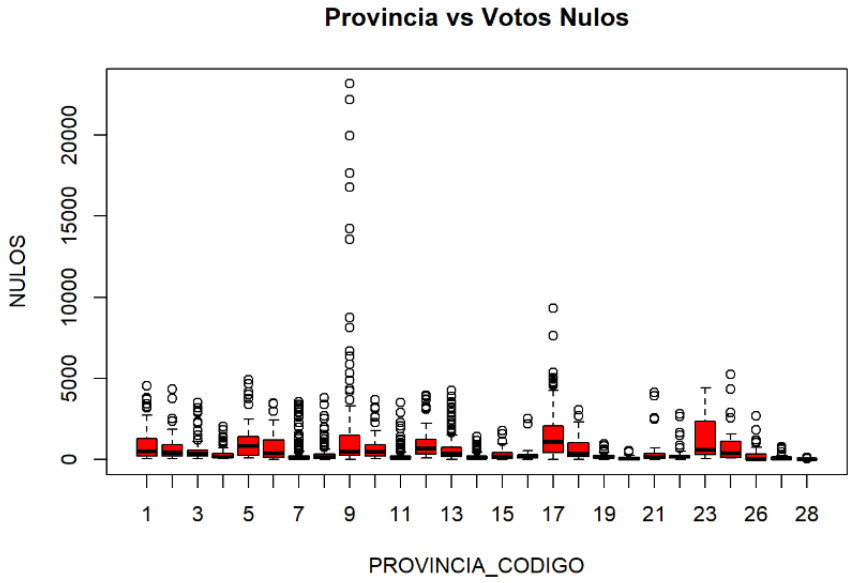

Figura 13. Gráfica votos nulos según la provincia

Otra de las gráficas que realizamos fue entre las provincias y los votos blancos (figura 14). La cual nos permite visualizar que existe un mayor número de votos blancos en la provincia con código 9 (Guayas).

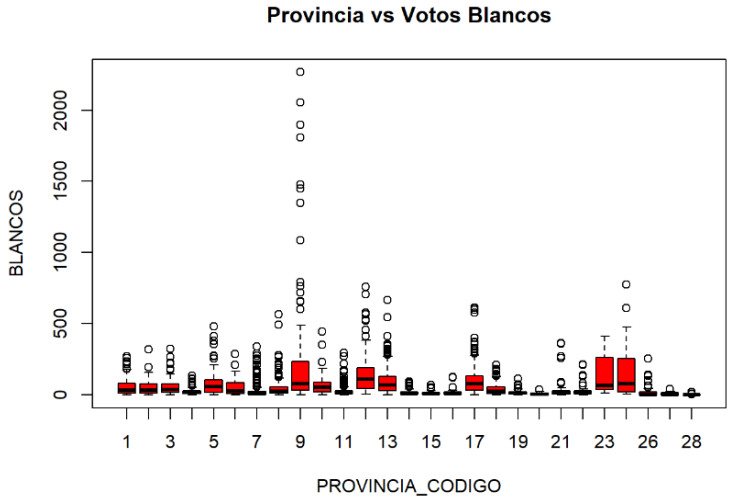

Figura 14. Gráfica votos nulos según la provincia

Otra de las gráficas que realizamos fue entre las provincias y los votos válidos, en donde pudimos visualizar que existe un mayor número de votos válidos en la provincia con código 9 (Guayas).

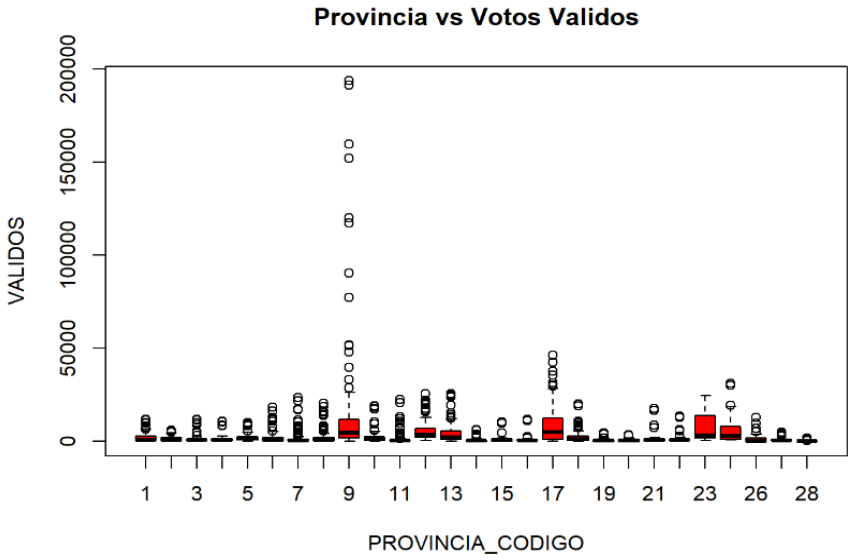

Figura 15. Gráfica votos nulos según la provincia

Entonces, de las gráficas presentadas anteriormente pudimos evidenciar que en la provincia de Guayaquil existe el mayor número de votos blancos, nulos y válidos.

Por otra parte, para el analisis que realizamos a los datos del INEC, generamos las siguientes graficas.

En la figura 16, nos muestra una grafica del número de personas alfabetas según la provincia; y pudimos observar que en la provincia 19 (Zamora Chinchipe) existe un mayor registro de personas alfabetas.

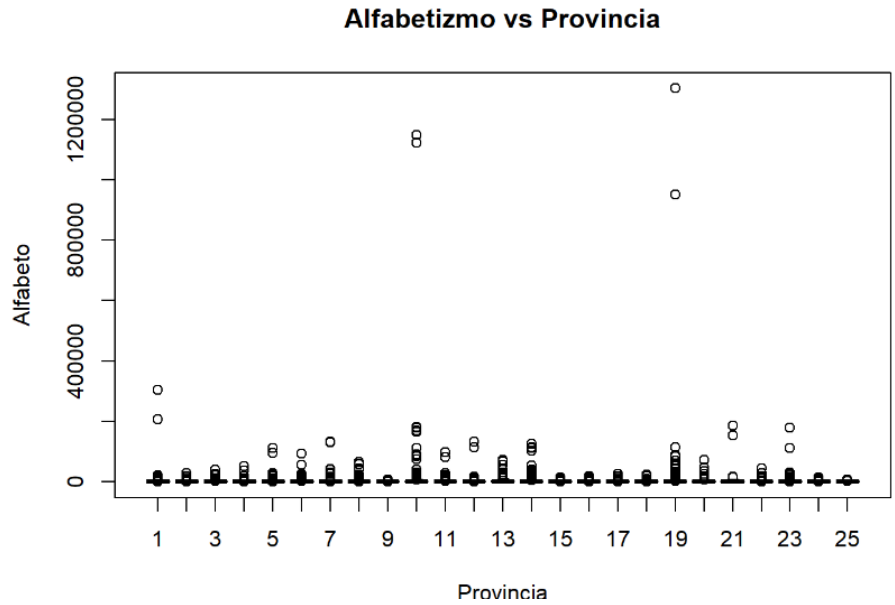

Figura 16. Gráfica de número de alfabetos según la provincias

De igual manera realizamos un gráfico entre el número de personas analfabetas según la provincia y pudimos observar que existe una mayor distribución en los datos en la provincia 10 (Imbabura).

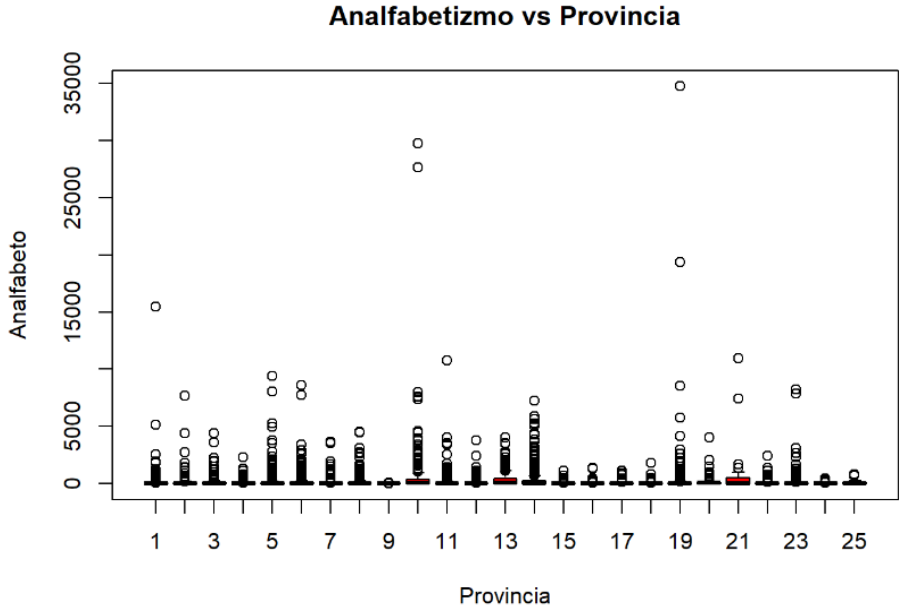

Figura 17. Gráfica de número de analfabetos según la provincias

Elaboración de un documento reproducible del análisis exploratorio

Parte de este trabajo incluye que realicemos de un documento reproducible con el cual, el lector pueda hacer seguimiento de todo el proceso que realizamos en el análisis exploratorio de los datos.

Dicho documento lo realizamos en RStudio con la ayuda del paquete Markdown y lo pueden encontrar el enlace en el anexo 3.

Observaciones finales

- En la base de datos del INEC se encontraron 42.7% de registros vacíos. De los cuales el 99.81% corresponden a provincia y el 98.5% a cantón. Esto debido a la estructuración de datos. En la sección 2.1 se realizó un proceso de limpieza y depuración para estos datos.
- Según datos del INEC existe 0.06% de personas analfabetas. El 18.87% pertenece a la provincia de Imbabura y el 14.06% a Morona Santiago. Por otro lado, el 0.94% restante son personas alfabetos. De ese porcentaje, el 26.09% pertenece a la provincia de Imbabura y el 19.25% a Zamora Chinchipe.
- De los datos proporcionados por el CNE las provincias de Pichincha y Guayas poseen los porcentajes más altos de votos blancos, nulos y válidos. Del 0.02% de votos blancos totales, el 20.09% pertenece a la provincia de Guayas y el 11.99% a Pichincha. Mientras que del 0.16% de votos nulos totales, el 17.18% pertenece a la provincia de Pichincha y el 16.43% a Guayas. Finalmente, del 0.82% de votos válidos totales, el 25.89% pertenece a la provincia de Guayas y el 17.92% a Pichincha.
- Hemos hallado un coeficiente de relación entre el analfabetismo y los votos blancos de -0.11, un -0.14 con los votos nulos y un -0.19 con los votos válidos.

3.1.2 Resultado análisis computacional integral

En esta sección presentamos los resultados que obtuvimos al finalizar el análisis computacional integral. Los resultados los mostraremos por cada nivel que definimos en la sección de metodología: arquitectura, sistema operativo, ambiente de programación, seguridad de la información e investigación.

3.1.2.1 Resultados Arquitectura

Procesamiento: Al analizar las dos arquitecturas encontramos que:

- La arquitectura ARM es ideal para tareas simples del análisis exploratorio como los cálculos aritméticos. Además, su consumo de energía es bajo. Sin embargo, está orientada a sistemas embebidos y dispositivos móviles.
- Por otro lado, la arquitectura x86, es más compleja para las tareas del análisis exploratorio, pero es accesible en el mercado porque está orientado a ordenadores. Además, su consumo de energía es mayor.

Almacenamiento: Al analizar las dos formas de almacenamiento encontramos que:

- Los discos SSD son superiores a los HDD debido a su bajo consumo de energía y mayor velocidad de arranque. Por otro lado, HDD es superior en cuanto a la vida útil.
- Además, debido a que R demanda que los datos este en memoria RAM para las tareas de análisis exploratorio de datos. Una buena opción sería la memoria DDR SDRAM que puede realizar dos instrucciones le lectura – escritura por cada ciclo de reloj.

Entrada – Salida: Al analizar las dos formas de entrada - salida encontramos que:

- Para la obtención de datos se necesita un sistema de interconexión que evite el paso por memoria RAM. Es decir, necesita un acceso directo a cache.
- Además, encontramos que la velocidad que nos ofrece PCIe frente a Sata es mayor. Esto nos beneficia ya que R demanda que los datos estén en memoria para utilizarlos.

3.1.2.2 Resultados Sistema Operativo

Procesamiento: R no se puede beneficiar directamente de procesadores multinúcleo, pero si se organizan las tareas estadísticas del análisis exploratorio de manera paralela, lo mejor sería utilizar hilos. Esto debido a su bajo consumo en memoria y CPU.

Almacenamiento: Al tener los datos en memoria los tiempos de acceso a almacenamiento se disminuyen. Sin embargo, esta limitado a las capacidades físicas de la memoria RAM. Además, debido a que R mantiene los datos en RAM, es necesario guardar los resultados del análisis exploratorio de datos en disco.

Entrada – Salida: El TIC utiliza pocos archivos y, potencialmente, de gran tamaño. Por lo tanto, el sistema de archivos debe ser configurado para satisfacer dichos requerimientos. Se debe crear una partición que esté montada directamente en el directorio raíz; por ejemplo: "/Datos", con una estructura jerárquica mínima, para minimizar el tiempo de apertura de los archivos de datos. Por otro lado, el sistema de archivos debe ser configurado con tamaños de bloques superiores a los tamaños de la configuración por defecto que utilizan los sistemas operativos para minimizar el tráfico entre memoria interna y externa.

Si trabajos similares al nuestro demandan el procesamiento de datos masivos, se debe tomar en cuenta el uso de sistemas de archivos distribuidos como HDFS de Hadoop, los cuales están diseñamos para manejar grandes volúmenes de datos.

Finalmente, después de realizar el análisis integral encontramos que las tareas del análisis exploratorio necesitan más recurso computacional de los siguientes componentes: CPU, RAM, disco y tarjeta de red.

3.1.2.3 Resultados Ambiente de programación

Procesamiento: R cuenta con bibliotecas estadísticas para diferentes aplicaciones; tales como: dplyr, haven, splines, parallel, nlme, nnet, etc.

Esto facilita el desarrollo de las aplicaciones estadísticas, pero, como toda biblioteca, las funciones son genéricas, y, por lo tanto, pueden ser ineficientes.

Almacenamiento: Las matrices ocupan menor espacio de memoria, pero no permite una visualización estructurada de los datos. Sin embargo, los data frame permite una visualización como una hoja de datos, pero consume más espacio de memoria.

Entrada – Salida: Los resultados se almacenan temporalmente en la memoria RAM y para presentar los resultados utiliza las bibliotecas de R, así como de los controladores gráficos de cada computador.

3.1.3 Resultado evaluación de rendimiento

En esta sección presentamos los resultados que obtuvimos al finalizar la evaluación de rendimiento. Los resultados serán presentados con base en el procesamiento, almacenamiento y entrada salida de cada infraestructura.

Después de finalizar con la ejecución de las tareas del análisis exploratorio en cada infraestructura, obtuvimos los siguientes resultados:

Infraestructura Local

En la gráfica 18 mostramos el porcentaje de utilización del procesador en las tareas del análisis exploratorio. Podemos observar que existe un pico alto de consumo al comienzo del análisis. Eso nos dice que la tarea de limpieza de base de datos de INEC consume más recursos que las demás tareas. Además, también podemos evidenciar que el consumo máximo de CPU es del 71.65% para realizar las tareas del análisis exploratorio de datos.

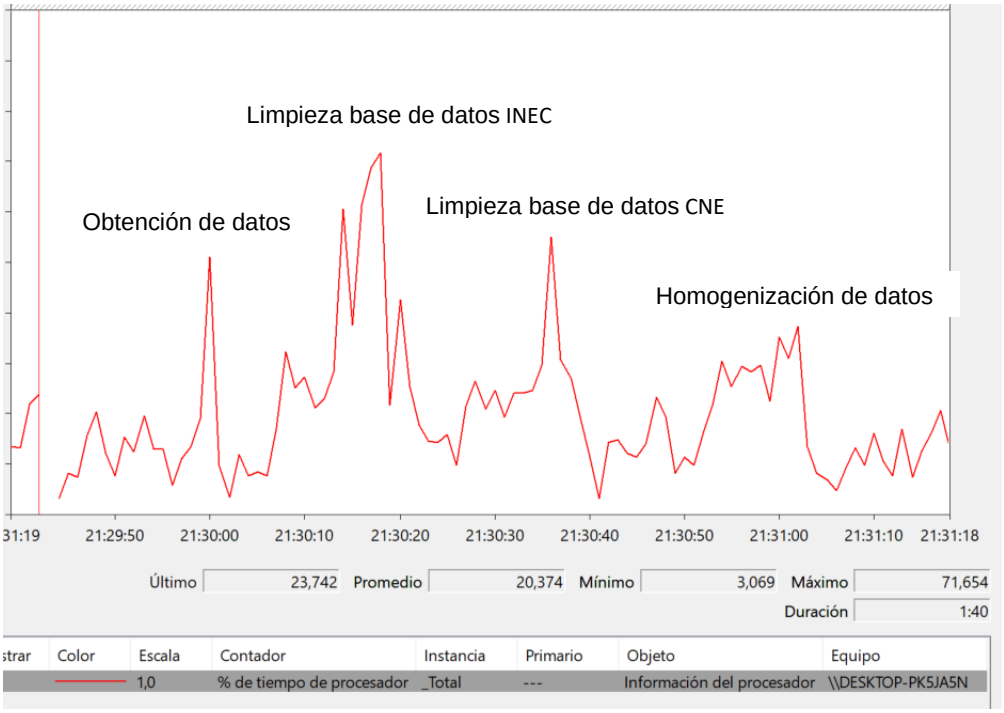

Figura 18. Consumo de CPU

En la figura 19 mostramos una gráfica de los paquetes y bytes enviados y recibidos por el adaptador de red. Podemos observar que existen varios picos. Estos picos sucedieron al momento de ejecutar la tarea de recolección de datos de la web. Entonces podemos decir que para ejecutar la tarea mencionada necesitamos más recursos computacionales de entrada – salida.

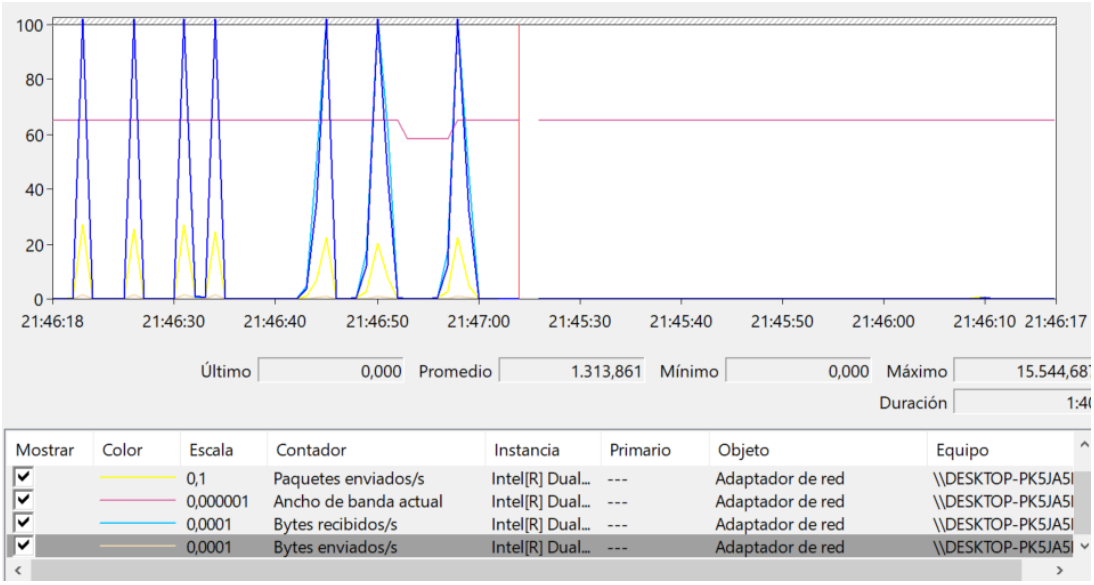

Figura 19. Paquetes y bytes enviados - recibidos por el adaptador de red

Por último, en esta figura 20 mostramos el porcentaje de lectura y escritura a disco. Podemos observar la existencia de un pico alto de consumo. Esto sucedió al momento de guardar los resultados del análisis exploratorio de datos.

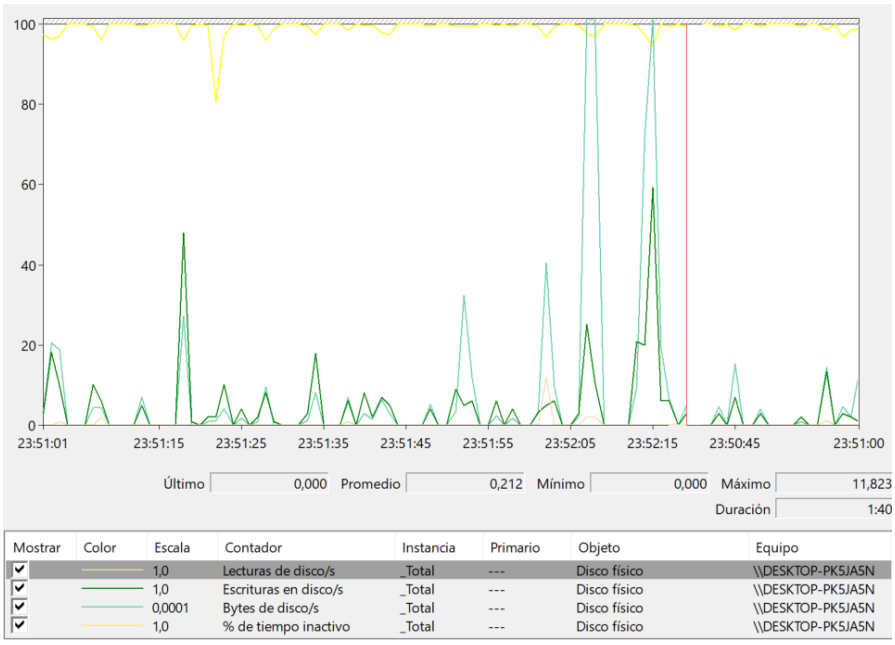

Figura 20. Porcentaje de lectura y escritura a disco.

Infraestructura en la nube

Después de realizar las tareas en la infraestructura en la nube obtuvimos los siguientes resultados:

En la figura 21, podemos observar que el porcentaje de uso de la CPU es mayor al iniciar con las tareas del análisis exploratorio. Sin embargo, después el consumo disminuye a porcentajes bajos, esto debido a que, en la tarea de obtención de datos, estos llegan a atreves de la red y deben ser cargados en memoria. Lo que causa un consumo considerable de CPU. Además, podemos observar que el consumo máximo de CPU es del 30%.

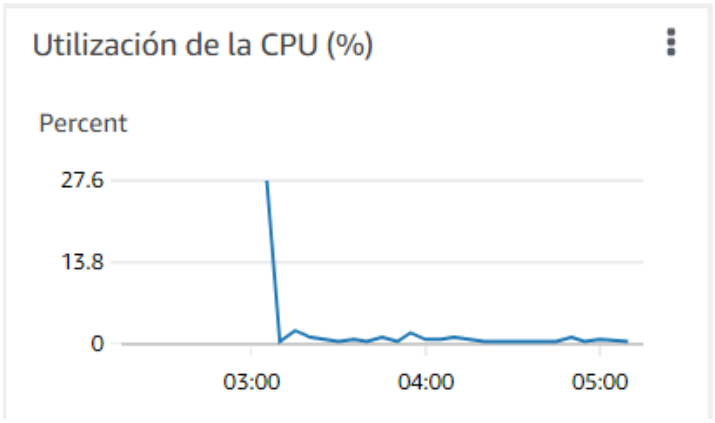

Figura 21. Utilización de CPU en AWS.

En la figura 22 podemos observar que los bytes de entrada de red tienen dos picos. Lo cual sucede en el momento de realizar la tarea de obtención de datos desde la web. En la otra grafica de salida de red, podemos observar que solo existe un pico de bytes de salida al comenzar con la tarea de obtención de datos.

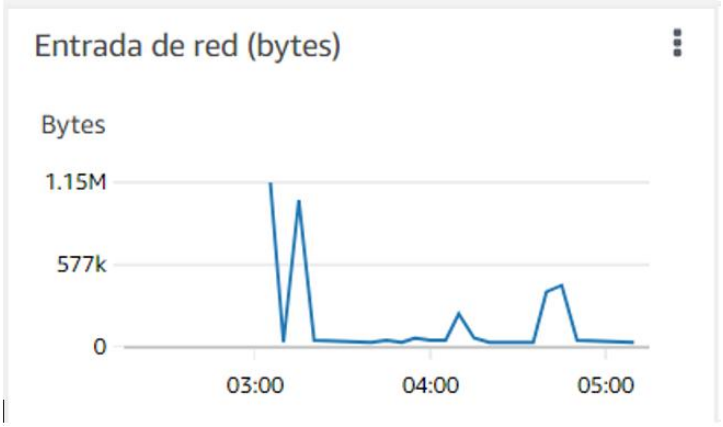

Figura 22. Bytes de entrada de red en AWS.

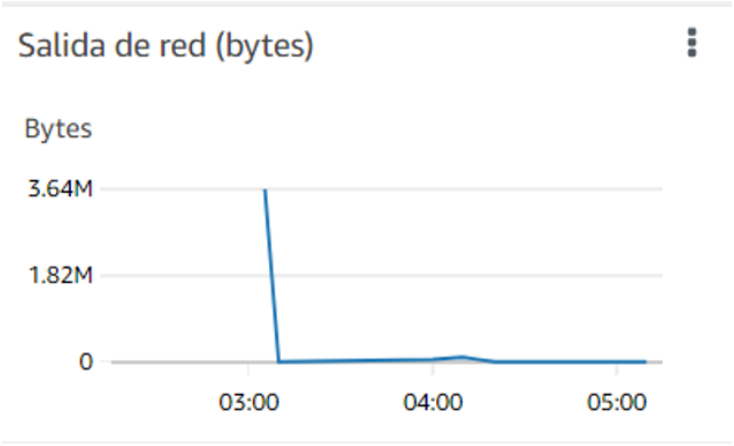

Figura 23. Bytes de salida de red en AWS.

En las figuras 24 y 25 podemos observar que los bytes de lecturas y escrituras de disco se mantienen en porcentajes bajos. Esto sucedió en todo el proceso de ejecución de las tareas del análisis exploratorio de datos.

Figura 24. Bytes de lectura de disco en AWS.

Figura 25. Bytes de escritura de disco en AWS.

3.2 Conclusiones

En esta sección de documento se detallan las conclusiones halladas. Las conclusiones están divididas en 3 partes como se mencionó en el alcance del documento. En la primera parte están las conclusiones del análisis exploratorio de datos. En la segunda parte, las conclusiones del análisis computacional integral. Por último, se detallan las conclusiones de la evaluación de rendimiento.

Análisis exploratorio de datos

Del análisis exploratorio de datos realizado hemos llegado a concluir que existe una relación estadísticamente significativa entre el nivel de analfabetismo y los resultados electorales. Hemos hallado un coeficiente de relación entre el analfabetismo y los votos blancos de -0.11, un -0.14 con los votos nulos y un -0.19 con los votos válidos. Es decir, mientras el total de votos válidos aumenta el porcentaje de personas analfabetas se reduce significativamente. Lo mismo sucede con los votos nulos y blancos. También encontramos una relación estadística entre los mismos tipos de votos. Encontramos un coeficiente de 0.91 entre los votos blancos y nulos. Eso nos dice que el total de votos blancos crece a la par del total de votos nulos.

Análisis computacional integral

En base a los resultados obtenidos hemos determinado que para la obtención de datos necesitamos ampliar nuestro ancho de banda en el caso de que el TIC requiera utilizar grandes volúmenes de datos. Así como utilizar un sistema de interconexión que evite el paso por memoria RAM, es decir un acceso directo a cache. Además, para la tarea de limpieza y depuración de datos necesitamos una memoria RAM con al menos 16 GB de capacidad. Esto debido a que R demanda que los datos estén cargados en memoria. Por otra parte, para reducir los tiempos de acceso a almacenamiento el sistema de archivos debe ser configurado con tamaños de bloques superiores a los tamaños de la configuración por defecto que utilizan los sistemas operativos. Con ello podemos minimizar el tráfico entre memoria interna y externa. Finalmente encontramos que las tareas del análisis exploratorio necesitan más recurso computacional de los siguientes componentes: CPU, RAM, disco y tarjeta de red.

Evaluación de rendimiento

Al concluir con la evaluación de rendimiento hemos determinado que el consumo de recursos computacionales en la infraestructura local fue del 71.65%. A diferencia del

consumo en una infraestructura en la nube que fue del 30%. Además, a pesar del alto porcentaje de uso computacional en la infraestructura local, se llevó a cabo todas las tareas del análisis exploratorio con éxito. Sin embargo, como se mencionó anteriormente se evidencio un alto consumo de recursos de memoria y CPU al ejecutar las tareas. Mientras que, en la infraestructura en la nube, se evidencio un consumo bajo de CPU al igual que el total de bytes de entrada salida de red.

3.3 Recomendaciones

En esta sección del documento se presentan las recomendaciones para mejoras y trabajos futuros de este proyecto. Estos están basados en las 3 partes de desarrollo del análisis como se mencionó en el alcance del documento. En la primera parte se detallan las conclusiones del análisis exploratorio de datos. En la segunda parte, las conclusiones del análisis computacional integral. Por último, se detallan las conclusiones de la evaluación de rendimiento.

Análisis exploratorio de datos

En base a los resultados encontrados en el análisis exploratorio de datos recomendamos como investigación futura realizar un análisis de las siguientes elecciones presidenciales con el censo poblacional que se va a realizar en este 2022. Con ello se puede profundizar en la correlación entre los resultados electorales y las variables socioeconómicas en Ecuador. Como investigación alternativa se puede analizar los resultados electorales y los niveles de empleo y desempleo en Ecuador.

Análisis computacional integral

Para este análisis recomendamos realizar las siguientes modificaciones: utilizar mayor capacidad de memoria RAM para reducir la limitación por el tamaño de los datos, configurar un tamaño de bloque mayor al utilizado por defecto y con ello poder evitar el tráfico de acceso a memoria interna y externa, además también se puede considerar almacenar los datos y resultados en memoria una externa. Con estos cambios se puede realizar un nuevo análisis computacional integral utilizando las tareas del análisis exploratorio de datos y analizar si los cambios significaron un cambio al realizar dichas tareas.

Evaluación de rendimiento

Al finalizar con la evaluación de rendimiento recomendamos realizar una planificación de los requerimientos computacionales necesarios para realizar un análisis de datos. Con ello se puede volver a realizar una nueva evaluación de rendimiento y analizar si los requerimientos definidos en la planificación satisficieron las necesidades computacionales; tanto para las tareas del análisis, así como de futuros análisis.

4 REFERENCIAS BIBLIOGRÁFICAS

[1] "Investigación Reproducible y Análisis de Datos · GitBook." [Online]. Available: https://open-science-training-handbook.github.io/Open-Science-Training-Handbook_ES/02OpenScienceBasics/04ReproducibleResearchAndDataAnalysis.html. [Accessed: Jul. 10, 2022]

[2] CC2020 Task Force, *Computing Curricula 2020: Paradigms for Global Computing Education*. New York, NY, USA: ACM, 2020 [Online]. Available: https://dl.acm.org/doi/book/10.1145/3467967. [Accessed: Aug. 11, 2022]

[3] "malla_computación.pdf." [Online]. Available: https://atenea.epn.edu.ec/bitstream/25000/604/3/malla_computaci%c3%b3n.pdf. [Accessed: Aug. 11, 2022]

[4] G. Westreicher, "Distribución de frecuencias," *Economipedia*. [Online]. Available: https://economipedia.com/definiciones/distribucion-de-frecuencias.html. [Accessed: Aug. 11, 2022]

[5] J. F. López, "Medidas de dispersión," *Economipedia*. [Online]. Available: https://economipedia.com/definiciones/medidas-de-dispersion.html. [Accessed: Aug. 11, 2022]

[6] "Coeficiente de correlación de Pearson: qué es y cómo se usa." [Online]. Available: https://psicologiaymente.com/miscelanea/coeficiente-correlacion-pearson. [Accessed: Jul. 28, 2022]

[7] "(PDF) Economía, política social y Twitter: análisis de las emociones negativas en cuatro elecciones presidenciales latinoamericanas a través del LIWC." [Online]. Available: https://www.researchgate.net/publication/339090895_Economia_politica_social_y_Twitter_analisis_de_las_emociones_negativas_en_cuatro_elecciones_presidenciales_latinoamericanas_a_traves_del_LIWC. [Accessed: Jan. 06, 2022]

[8] A. Panchana-Macay and C. Barrera, "Ecuador TV como medio de propaganda en las elecciones presidenciales de la era Correa (2007-2017)," *Revista de Comunicación*, vol. 20, no. 2, pp. 319–337, Sep. 2021, doi: 10.26441/rc20.2-2021-a17.

[9] "Estructura de computadores." [Online]. Available: http://cv.uoc.edu/annotation/8255a8c320f60c2bfd6c9f2ce11b2e7f/619469/PID_00218274/PID_00218274.html. [Accessed: Aug. 12, 2022]

[10] D. S. Llaven, *Sistemas Operativos: Panorama para ingeniería en computación e informática*. Grupo Editorial Patria, 2015.

[11] U. Ruelas, "¿Qué es la computación concurrente?" [Online]. Available: https://codingornot.com/que-es-la-computacion-concurrente. [Accessed: Aug. 11, 2022]

[12] "La computación paralela: alta capacidad de procesamiento," *Teldat Blog - Connectando el mundo*, Jul. 14, 2020. [Online]. Available: https://www.teldat.com/blog/es/computacion-paralela-capacidad-procesamiento/. [Accessed: Aug. 11, 2022]

[13] U. Ruelas, "Computación distribuida." [Online]. Available: https://codingornot.com/computacion-distribuida. [Accessed: Aug. 11, 2022]

[14] "El concepto de IDE." [Online]. Available: https://www.redhat.com/es/topics/middleware/what-is-ide. [Accessed: Aug. 12, 2022]

[15] "Estructuras de Datos · ciencia-de-datos-con-r." [Online]. Available: https://rsanchezs.gitbooks.io/ciencia-de-datos-con-r/content/estructuras_datos/estructuras_datos.html. [Accessed: Aug. 12, 2022]

[16] "ITIL 4: PRÁCTICAS DE GESTIÓN DE ITIL: ADMINISTRACIÓN DE CAPACIDAD Y RENDIMIENTO," *Interpolados*, Sep. 22, 2020. [Online]. Available: https://interpolados.wordpress.com/2020/09/22/itil-4-practicas-de-gestion-de-itil-administracion-de-capacidad-y-rendimiento/. [Accessed: Aug. 12, 2022]

[17] Rafa, "Análisis exploratorio de datos (o EDA) con R," *Rafa González Gouveia*, May 16, 2020. [Online]. Available: https://gonzalezgouveia.com/analisis-exploratorio-de-datos-en-r/. [Accessed: Jul. 02, 2022]

5 ANEXOS

ANEXO I. Lista de códigos

Lista de códigos de provincia y Etnia

Código	Nombre de Provincia
1	Azuay
2	Bolívar
3	Cañar
4	Carchi
5	Cotopaxi
6	Chimborazo
7	El Oro
8	Esmeraldas
9	Guayas
10	Imbabura
11	Loja
12	Los Ríos
13	Manabí
14	Morona Santiago
15	Napo
16	Pastaza
17	Pichincha
18	Tungurahua
19	Zamora Chinchipe
20	Galápagos
21	Sucumbíos
22	Orellana
23	Sto. Dgo. Tsáchilas
24	Santa Elena
25	Zonas no delimitadas
26	Europa Asia y Oceanía
27	EE. UU y Canadá
28	América latina y El Caribe

Código	Etnia
1	Afroecuatoriano/a
2	Blanco/a
3	Indígena
4	Mestizo/a
5	Montubio/a
6	Otro/a

Anexo II. Enlace documento reproducible

Enlace del documento reproducible:

Documento Reproducible.html