

**UNIVERSIDAD TECNOLÓGICA DE PEREIRA**

**FACULTAD DE INGENIERÍAS**

**MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN**



**Universidad  
Tecnológica  
de Pereira**

**IMPLEMENTACIÓN DE UN MODELO ANALÍTICO DE DATOS EXPLORATORIO Y  
ESTADÍSTICO DEL BILINGÜISMO EN LOS RESULTADOS DE LAS PRUEBAS  
SABER 11 DE LAS INSTITUCIONES EDUCATIVAS A NIVEL NACIONAL  
EN EL PERIODO 2015 AL 2019.**

**TESIS DE MAESTRÍA**

**DIRIGIDA POR:**

**CARLOS ANDRÉS LÓPEZ**

**PRESENTADA POR:**

**VALENTINA GUEVARA SERNA**

**LUIS FERNANDO ZULUAGA JARAMILLO**

## **DEDICATORIA**

Dedicado, en agradecimiento a Dios, porque sin Él nada de esto podría ser posible. A nuestros padres y familiares, por su apoyo permanente en cada etapa de nuestra vida.

A Msc. Carlos Andrés López, por su gran profesionalismo y valioso acompañamiento para culminar esta investigación.

## **AGRADECIMIENTOS**

Le agradecemos a nuestro director de tesis Msc. Carlos Andrés López, quien gracias a su gran profesionalismo nos pudo acompañar desde los inicios del proceso y nos motivó constantemente hasta el final de esta valiosa investigación.

Igualmente, a cada uno de los profesores que nos acompañaron en cada una de las asignaturas como pilares fundamentales que con su profesionalismo nos brindaron el conocimiento necesario para culminar la investigación.

## TABLA DE CONTENIDO

CAPITULO 1: DEFINICIÓN DEL PROBLEMA .....	9
1.1 Descripción del Problema .....	9
1.2 Formular el problema .....	11
1.3 Objetivos de la investigación .....	12
1.4 Justificación de la investigación.....	13
1.5 Viabilidad y alcance de la investigación.....	17
1.6 Metodología .....	18
1.6.1 Hipótesis .....	18
1.6.2 Diseñar modelo en espiral de prototipo de flujo de trabajo exploratorio. ....	18
CAPÍTULO II: MARCO TEÓRICO .....	20
2.1 Fundamentación teórica .....	20
2.1.1 Modelo en espiral de prototipo de flujo.....	20
2.1.2 Ciencia de Datos.....	21
2.1.3 Extract, Transform and Load ETL .....	29
2.1.4 Estadística y analítica de datos .....	34
CAPÍTULO III: DESARROLLO DE LA TESIS.....	47
3.1 Recolectar datos de los resultados de las pruebas Saber 11 .....	47
3.2 Realizar el proceso de Extracción, Transformación y Carga (ETL) sobre los datos recolectados de las pruebas Saber 11 .....	49

3.3 Implementar un lago de datos para gestionar los datos crudos pre procesados y procesados .....	55
3.4 Identificar las variables de valor sobre los datos a procesar .....	56
3.5 Buscar patrones de incidencia y valores estadísticos sobre las variables identificadas .....	59
3.6 Cargar datos en modelo de análisis multivariado.....	68
3.7 Crear el modelo descriptivo utilizando analítica de datos sobre las variables de valor .....	71
3.8 Validar los patrones hallados mediante pruebas estadísticas .....	76
3.9 Analítica diagnóstica de datos sobre las variables de valor .....	79
<b>CAPÍTULO IV: CONCLUSIONES Y RECOMENDACIONES .....</b>	<b>81</b>
4.1 Conclusiones .....	81
4.2 Recomendaciones y Trabajos futuros .....	83
<b>CAPÍTULO V: REFERENCIA BIBLIOGRÁFICA .....</b>	<b>84</b>

## LISTA DE TABLAS

Tabla 1 Escalas de los puntajes en el examen Saber 11 a partir de 2014-II. ....	14
Tabla 2 Rango niveles de desempeño Puntaje Global Pruebas Saber 11 .....	57
Tabla 3 Rango niveles de desempeño Puntaje Inglés Pruebas Saber 11 .....	58
Tabla 4 Número de estudiantes por población privado - público bilingüe .....	59
Tabla 5 Agrupación de puntuaciones Lectura crítica .....	60
Tabla 6 Agrupación de puntuaciones Matemáticas .....	61
Tabla 7 Agrupación de puntuaciones Ciencias Naturales.....	61
Tabla 8 Agrupación de puntuaciones Ciencias Sociales.....	62
Tabla 9 Agrupación de puntuaciones inglés .....	62
Tabla 10 Agrupación de puntuaciones Global.....	63
Tabla 11 Porcentaje de puntuaciones Global por Colegios .....	64
Tabla 12 Promedio datos atípicos de colegios públicos .....	65
Tabla 13 Promedio datos atípicos de colegios privados bilingües .....	66
Tabla 14 Promedio datos atípicos de colegios privados no bilingües .....	67
Tabla 15 Porcentaje comparativo de puntuaciones Global por Colegios .....	70
Tabla 16 Naturaleza del Colegio.....	71
Tabla 17 Colegios Bilingües .....	71
Tabla 18 Colegios Bilingües Vs Naturaleza del Colegio.....	72
Tabla 19 Niveles de Desempeño Puntaje Global Pruebas Saber 11 .....	75
Tabla 20 Porcentaje de estudiantes por Puntaje Global Pruebas Saber 11 .....	76

## LISTA DE FIGURAS

Ilustración 1 Cálculo del Puntaje Global .....	14
Ilustración 2 Modelo de Espiral .....	21
Ilustración 3 Estructura bases de datos SQL y NoSQL .....	26
Ilustración 4 Teorema CAP .....	27
Ilustración 5 Logo Python.....	30
Ilustración 6 Logo R .....	31
Ilustración 7 Nivel de madurez de los datos valor y complejidad .....	39
Ilustración 8 Madurez de datos decisión y acción .....	39
Ilustración 9 Caja de bigotes.....	44
Ilustración 10 Histograma .....	45
Ilustración 11 Proceso de transformación de separadores de datos .csv.....	51
Ilustración 12 Panel de Control Instancias AWS.....	52
Ilustración 13 Información de db Icfes en Mongo.....	52
Ilustración 14 Exportar de db Icfes con filtros de Mongo .....	53
Ilustración 15 Cálculo del Puntaje Global .....	53
Ilustración 16 Librerías usadas en Rstudio .....	59
Ilustración 17 Bigotes privado puntaje global .....	64
Ilustración 18 Bigotes público puntaje global .....	64
Ilustración 19 Carga y filtro de datos en R Studio.....	68
Ilustración 20 Caja de Bigotes en R Studio .....	69
Ilustración 21 Caja de Bigotes Puntaje Global Oficial vs No Oficial.....	72
Ilustración 22 Histogramas Puntaje Global publico .....	73

Ilustración 23 Histogramas Puntaje Global Privado no bilingüe.....	74
Ilustración 24 Histogramas Puntaje Global Privado Bilingüe .....	74
Ilustración 25 Prueba de hipótesis 1 en RStudio privado .....	77
Ilustración 26 Prueba de hipótesis 1 en RStudio publico .....	77
Ilustración 27 Prueba de hipótesis 2 en RStudio privado .....	78
Ilustración 28 Prueba de hipótesis 2 en RStudio publico .....	78

# CAPITULO 1: DEFINICIÓN DEL PROBLEMA

## 1.1 Descripción del Problema

La educación es un factor trascendental en el desarrollo de un país, puesto que un talento humano bien educado y con buenas competencias académicas contribuye a avanzar en todas las áreas en las que se encuentre, sea economía, tecnología, salud, gobierno entre otras. Por tal motivo la educación en Colombia busca constantemente mejorar el nivel educativo y ser competitivo a nivel internacional, así al empezar a participar en las pruebas PISA, se identifica a nivel estatal la gran brecha que existe en comparación con el nivel alcanzado y observado en otros países, por tal motivo es de suma importancia identificar algún factor diferencial en las pruebas Saber 11 [1], que pueda determinar los focos de las acciones que se están implementando de mejor manera y desde allí poderlas considerar como experiencias positivas, tratando de generar una mejora continua.

Al mirar los resultados se encontró que no estamos en un buen nivel, solo el 5% de los mejores estudiantes colombianos en matemáticas tiene el promedio de los estudiantes de Finlandia [3], visualizando así una brecha enorme entre Colombia y otros países. Entonces, ¿qué estamos haciendo mal?, ¿qué debemos mejorar?, ¿hacia dónde debemos enfocarnos para mejorar la educación en Colombia? Aunque en algunos de los estudios como el de Luz Karime Abadía Alvarado, Gloria Lucía Bernal Nisperuza y Santiago Muñoz González quienes estudiaron las brechas en el desempeño escolar en PISA a través del artículo: ¿que explica la diferencia de Colombia con Finlandia y Chile? [3], lograron identificar que, al analizar factores personales, familiares y algunos escolares, dichos factores no daban una explicación clara de la brecha.

Teniendo en cuenta el ranking de los mejores colegios en las pruebas Saber 11 del 2019 en Colombia [4], 7 de los 10 mejores colegios son Bilingües y todos son privados, este resultado genera un elemento determinante e importante para tener en cuenta para iniciar una investigación que arroje resultados cuantitativos de si realmente este factor impacta en los resultados de las pruebas Saber 11 de los estudiantes de Colombia.

Uno de estos factores determinantes ha sido influenciado por el fenómeno acelerado de la globalización; el bilingüismo ha sido introducido en algunas instituciones como referente comercial y de competencia frente a sus similares, pero no ha sido un proceso sencillo de incorporar en los planes académicos de ningún colegio, sin embargo, en la actualidad posiciona en los primeros lugares de resultados de las pruebas Saber 11 a aquellos que han avanzado lo suficiente en este osado proceso [5].

## **1.2 Formular el problema**

¿Se puede explorar mediante métodos estadísticos, la existencia de vínculos entre el bilingüismo y el desempeño global de las instituciones educativas en las pruebas Saber 11, utilizando la ciencia de datos?

## **1.3 Objetivos de la investigación**

### **Objetivo general**

Utilizar la ciencia de datos para explorar si existe un vínculo estadístico entre el bilingüismo y los resultados globales de las pruebas Saber 11.

### **Objetivos específicos**

- Recolectar datos de los resultados de las pruebas Saber 11.
- Realizar el proceso de Extracción, Transformación y Carga (ETL) sobre los datos recolectados de las pruebas Saber 11.
- Implementar un lago de datos para gestionar los datos crudos pre procesados y procesados.
- Identificar las variables de valor sobre los datos procesados.
- Calcular valores estadísticos sobre las variables identificadas (Big Data e inferencia estadística).
- Crear el modelo descriptivo utilizando analítica de datos sobre las variables de valor.
- Buscar patrones de incidencia sobre las variables identificadas.
- Validar los patrones hallados mediante pruebas estadísticas.

## **1.4 Justificación de la investigación**

### **Antecedentes**

Desde 1968 el ICFES (Instituto Colombiano para el Fomento de la Educación Superior) se enfoca principalmente en evaluar las habilidades de los estudiantes de Colombia, con el objetivo inicial de filtrar el ingreso a la educación superior y posteriormente mejorar la calidad de la educación nacional [1]. Por esto, en Colombia existe la evaluación Saber 11 estandarizada y semestral a cargo del ICFES y aplicada a los estudiantes de educación media con el objetivo de: seleccionar estudiantes para la educación superior, monitorear la calidad de la formación de los estudiantes y una estimación del valor agregado de la educación superior. [31]

Con el objetivo de consolidar un Sistema Nacional de Evaluación Estandarizada (SNEE) que consiga la alineación de todos los exámenes que lo conforman, la estructura del examen Saber 11 fue modificada a partir del segundo semestre de 2014 para que sus resultados fueran comparables, en términos de la evaluación de competencias genéricas, con los de otras pruebas del SNEE como las pruebas Saber 3°, 5° y 9° realizadas a estudiantes de 3ro, 5to y 9no grado de escolaridad y el examen Saber Pro realizado a los estudiantes de educación superior. La estructura queda con las pruebas de Lectura crítica, Matemáticas (Incluye razonamiento Cuantitativo), Ciencias Naturales, Sociales y ciudadanas (Incluye Competencias Ciudadanas) e inglés.

Este examen se establece como metodología de calificación así:

$$PG = 5 * IG$$

Donde:

$$IG = \frac{3 * MATEMÁTICAS + 3 * LECTURA + 3 * CIENCIAS + 3 * SOCIALES + 1 * INGLÉS}{13}$$

*MATEMÁTICAS*: Puntaje en la prueba de matemáticas

*LECTURA*: Puntaje en la prueba de lectura

*CIENCIAS*: Puntaje en la prueba de ciencias naturales

*SOCIALES*: Puntaje en la prueba de sociales y ciudadanas

*INGLÉS*: Puntaje en la prueba de inglés

*Ilustración 1 Cálculo del Puntaje Global<sup>1</sup>*

Escalas de 2014-2 en adelante		
Puntaje	Mínimo	Máximo
Por Prueba	0	100
Global	0	500

*Tabla 1 Escalas de los puntajes en el examen Saber 11 a partir de 2014-II.*

Aunque esta prueba está enfocada para ser aplicada a estudiantes de educación media, existen 3 tipos de evaluados:

1. **Estudiantes**: evaluados que presentan el examen a través de una institución educativa y están en el último año de educación media.
2. **Validantes**: evaluados que presentan el examen para validar su bachillerato.
3. **Individuales**: evaluados que presentan el examen de forma individual y no a través de una institución educativa. En general, estos evaluados son estudiantes ya graduados.

<sup>1</sup> [https://www.icfes.gov.co/documents/20143/1885630/1.+Documentacion\\_Saber11.pdf/e72d7e45-7b05-fbee-aed7-c0dfafa25e2f?t=1590543922537](https://www.icfes.gov.co/documents/20143/1885630/1.+Documentacion_Saber11.pdf/e72d7e45-7b05-fbee-aed7-c0dfafa25e2f?t=1590543922537)

A pesar de los esfuerzos constante del ICFES para mejorar el modelo de evaluación, estandarizar los exámenes y así identificar falencias en el sector educativo encontrando datos significativos para mejorar la calidad de la educación en Colombia, esto no ha sido del todo posible ya que los resultados en las pruebas internacionales nos ha dejado con un gran margen a nivel educativo con respecto a otros países latinoamericanos como Chile y aún más si se compara con Finlandia, como en el caso de las pruebas PISA del año 2012. Se concluye que el promedio de los estudiantes de Finlandia es tan bueno como el promedio del 5% de los mejores estudiantes colombianos en matemáticas, pero el 89.37% de los estudiantes colombianos no superan un nivel básico en la misma área (OCDE niveles 1 y 2 de 6) [3].

### **Justificación**

Desde el inicio de la participación de Colombia en las pruebas PISA, se identifica de manera consistente la desigualdad que se presenta con respecto a otros países desarrollados, por ejemplo en el artículo “Brechas en el desempeño escolar en PISA: ¿Qué explica la diferencia de Colombia con Finlandia y Chile?”[3], se tienen en cuenta los datos arrojados por las pruebas PISA del año 2012 y se concluye que el promedio de los estudiantes de Finlandia es tan bueno como el promedio del 5% de los mejores estudiantes colombianos en matemáticas, pero el 89.37% de los estudiantes colombianos no superan un nivel básico en la misma área (OCDE niveles 1 y 2 de 6), mostrando que Colombia está lejos de alcanzar el nivel de los países desarrollados, y aunque se investigó sobre factores personales, familiares y escolares, el factor no explicable en Finlandia y Chile es del 80% y 41% respectivamente para esta brecha.

Pero la globalización ha obligado al estado y a las instituciones privadas a dirigir su mirada hacia el fenómeno del bilingüismo, aspecto que hace más de una década sólo era sinónimo de

educación para estratos altos y que 30 años atrás solamente aparecía dentro de las pruebas ICFES como una electiva, sin embargo, actualmente ha comenzado a ser parte activa dentro del conjunto de áreas educativas fundamentales establecidas por el MEN, junto al lenguaje, las ciencias y las matemáticas.

El bilingüismo aparte de ser un contenido específico para el desarrollo de habilidades en una segunda lengua podría comenzar a jugar un papel bastante significativo y relevante en el desempeño individual y global de un colegio en las pruebas de estado, no solo por el puntaje obtenido en esa modalidad sino por la habilidad neurológica y cognitiva que los estudiantes pueden desarrollar en otras asignaturas, la que influye en un mejor desempeño global aportado por todas las áreas. Por ello se hará un estudio cuantitativo y una implementación tecnológica, enfocados en el análisis de los resultados de las pruebas Saber 11 presentadas a nivel nacional en los últimos 5 años (2015-2019), tanto en instituciones de educación públicas como privadas, con el objetivo de determinar la diferencia que existe entre las instituciones bilingües y no bilingües, desde este ámbito.

La ciencia de datos tiene la capacidad de generar modelos exploratorios que utilizan métodos, procesos, algoritmos y sistemas científicos para extraer valor de los datos. El Científico de Datos combina habilidades, entre ellas estadística e informática para analizar datos recopilados.

## **1.5 Viabilidad y alcance de la investigación**

Actualmente el proyecto de investigación es viable, debido a que se cuenta con los medios y recursos de tipo económico y financiero, aspectos que serán asumidos por los estudiantes que lo proponen.

Existen los recursos materiales y el acceso a la documentación e información requerida para la implementación del estudio, porque las bases de datos necesarias para dicho propósito son de acceso público [5].

Además, hay disponibilidad del recurso humano y del tiempo necesario para el desarrollo del proyecto a nivel general, se cuenta con asesores expertos en el área y capacitación a través del posgrado en las asignaturas cursadas durante este, como Análisis Multivariado, Inferencia Estadística, Seminario I (servidores Cloud) y Big Data [6], entre otras habilidades mínimas necesarias.

## **1.6 Metodología**

### **1.6.1 Hipótesis**

Explorar mediante métodos estadísticos, la existencia de vínculos entre el bilingüismo y el desempeño global de las instituciones educativas en las pruebas Saber 11.

### **1.6.2 Diseñar modelo en espiral de prototipo de flujo de trabajo exploratorio.**

Se utilizar el modelo en espiral de prototipo de flujo se basará en un primer ciclo del modelo donde:

1. Se realiza la identificación e identificación del problema, objetivos y temática a tratar
2. Realizar la investigación relevante de antecedentes del problema e identificación de riesgos.
3. Como paso siguiente en la etapa de implementación se realiza:
  - a. Recolectar datos procedentes de datos abiertos de Colombia [5].
  - b. Extraer datos de los últimos 5 años de las pruebas Saber 11 para el procesamiento.
  - c. Crear un lago de datos [12] [13].
  - d. Preprocesar y limpiar datos crudos provenientes de archivos [7].
  - e. Cargar y unificar datos a bases de datos NoSQL [8] [9] [16].
  - f. Definir variables relevantes para el experimento.
  - g. Filtrar datos según definición de variables.
  - h. Identificar medidas de dispersión y análisis estadísticos de los datos [11].
  - i. Cargar datos en modelo de análisis multivariado [10].

- j. Explorar las clasificaciones obtenidas de modelos de aprendizaje de máquina [14] [15] [17] [18].
  - k. Visualizar el comportamiento de los datos en un modelo descriptivo [19].
  - l. Identificar patrones que vinculan el bilingüismo con los resultados generales de las pruebas.
4. Finalmente se efectúa la evaluación donde:
- a. Realizar prueba de hipótesis sobre hallazgos en modelos de aprendizaje de máquina.
  - b. Conclusiones.

## CAPÍTULO II: MARCO TEÓRICO

### 2.1 Fundamentación teórica

#### 2.1.1 Modelo en espiral de prototipo de flujo

Es un modelo que busca pasar controladamente y de manera cíclica por 4 actividades para el desarrollo de un experimento, los cuales son [41]:

- a. **Planificación:** Determinar los objetivos, alternativas de solución, las restricciones y encontrar el problema, así como aspectos administrativos de la investigación.
- b. **Análisis de riesgos:** Identificar los riesgos como la resolución. Estos riesgos pueden ser humanos, de conocimiento, económicos, culturales entre otros que puedan afectar el problema.
- c. **Ejecución:** es la investigación y ejecución de actividades enfocadas a la resolución del problema planteado.
- d. **Evaluación y conclusiones:** En esta etapa se genera una evaluación de la investigación, validando los resultados y realizando las conclusiones, las cuales se convierten en la materia prima del siguiente ciclo de la espiral.



*Ilustración 2 Modelo de Espiral<sup>2</sup>*

### 2.1.2 Ciencia de Datos

La ciencia de datos es hoy en día la principal herramienta para la explotación de los datos y a través de estos poder generar conocimiento. Entre los principales objetivos que persigue la ciencia de datos están la búsqueda de modelos que puedan describir patrones y comportamientos a partir de los datos y así poder tomar decisiones o quizás hacer predicciones. El área de la ciencia de datos ha experimentado un crecimiento amplio al incursionar en el acceso de grandes volúmenes de datos, que al mismo tiempo puedan ser tratados en tiempo real, lo que requiere de técnicas que puedan solucionar aspectos como la escalabilidad, robustez ante los errores y adaptabilidad con modelos dinámicos [19].

La ciencia de datos se ha convertido en materia de investigación de distintas áreas como la computación, la estadística, las matemáticas y la ingeniería, que trabajan en la propuesta de crear nuevos algoritmos, técnicas de computación e infraestructuras para la captura, almacenamiento y procesado de datos.

<sup>2</sup> [https://www.researchgate.net/figure/Figura-25-Ciclo-de-vida-en-espiral\\_fig2\\_39425588](https://www.researchgate.net/figure/Figura-25-Ciclo-de-vida-en-espiral_fig2_39425588)

Cuando encontramos que el manejo de los datos desborda o supera la capacidad de captura, almacenado, gestión y análisis de estos por medio de bases de datos convencionales, llegamos al punto de tener que procesar datos de manera masiva y es allí cuando comienza a jugar un papel bastante importante el término Big Data.

Big Data (grandes datos o macrodatos) se refiere a una multitud de tendencias tecnológicas nacientes desde la primera década del siglo XXI, donde se observó como en el año 2011 se crearon 1.8 zettabyte de datos (1 billón de gigabyte) y estos se duplican cada dos años según IDC (International Data Corporation). Aunque no se tiene una definición unánime de que es la big data, todas las definiciones se enfocan en el crecimiento exponencial de grandes volúmenes de datos, la captura, el almacenamiento y análisis de estos para conseguir el mayor beneficio y aprovechar las oportunidades que ofrecen.[39]

Big data en términos generales es información que no puede ser procesada o analizada usando procesos o herramientas tradicionales. Por tal motivo las técnicas de Big Data persiguen complementar el manejo de grandes volúmenes de datos con técnicas de análisis de la información más avanzadas y efectivas, para extraer de modo óptimo conocimiento del contenido de los datos, sin embargo, el concepto no solo hace referencia al tamaño de la información sino también a la variedad del contenido y a la velocidad con la que los datos se generan, almacenan y analizan. [40]

### **2.1.2.1 Servidor**

Un servidor como su nombre lo indica es una computadora o un conjunto de computadoras que está al servicio o prestan servicios a otras computadoras comúnmente nombradas clientes, los servidores son capaz de atender las peticiones de uno o muchos clientes y devolverle una

respuesta según la petición realizada. Un servidor puede estar alojado en cualquier tipo de computadora, aunque principalmente se utilizan computadores dedicados que se conocen como servidores.

Las bases de datos siguen un modelo de arquitectura conocido como cliente-servidor. Puesto que se tiene un programa central que accede directamente a los datos y ofrece servicios de información (consultas), edición o inserción que requieren los clientes.

Existen diferentes tipos de servidores desde uno local (en la misma computadora que se está trabajando) pasando por uno físico hasta los servidores en la nube o cloud, como por ejemplo Amazon AWS.

Un servidor cloud es un servidor virtual o un conjunto de máquinas virtuales que emulan el funcionamiento de un servidor físico. Los servidores cloud usualmente tienen o utilizan redundancia, por lo que la localización física exacta del servidor es difícil o hasta imposible de conocer. Esta redundancia hace que si un equipo cae (se apague, desconecte de la red o presente fallas para prestar servicios), el servidor cloud siga funcionando en otro equipo, manteniendo en todo momento, el funcionamiento y servicios alojados en dicho servidor.

Los proveedores de servidores cloud ofrecen servidores a través de una virtualización del servidor físico subyacente, que se distribuye de forma lógica en distintos servidores virtuales, cada uno con sus sistemas operativos, interfaz de usuario y aplicaciones.

En un servidor cloud podemos crear carpetas de datos vacías en las que podemos cargar la información que tenemos almacenada de manera local en algún formato estructurado de datos y sin importar si el servidor se inactiva, la información seguirá estando almacenada en este lugar [21].

### **2.1.2.2 Formato y estructura de los archivos de almacenamiento de los datos**

Para lograr analizar los datos que tenemos y convertirlos en información, debemos ser capaces de incorporarlos al programa o software que vamos a utilizar, esto hace parte de la fase de carga de datos. Los datos deben estar almacenados de una manera estructurada, de no ser así, procesar la información se convierte en una tarea muy difícil.

La forma de guardar la información debe tener una estructura fija, que permita de manera sencilla extraer los atributos de cada elemento, por eso, en el transcurso de los años han ido surgiendo distintos estándares para el almacenamiento de la información, entre estos están CSV, TSV, MS Excel, JSON, XML y SQL, para nuestro estudio hemos optado por el formato CSV.

Un archivo con extensión .csv es un documento con valores separados con comas, de ahí su nombre comma-separated values. En esencia constituye un sistema de organización de datos basado en columnas y filas, donde cada elemento constituye una fila o línea, en la que los atributos separados por comas deben siempre tener el mismo orden.

Por lo general, la primera fila llamada cabecera, no contiene datos de ningún tipo sino la información o estructura de los atributos de los elementos o registros.

El formato .csv se ha convertido en una estructura de almacenamiento de información que la mayoría de los programas puede leer y procesar sin ningún problema, entre ellos Python, R y MongoDB.

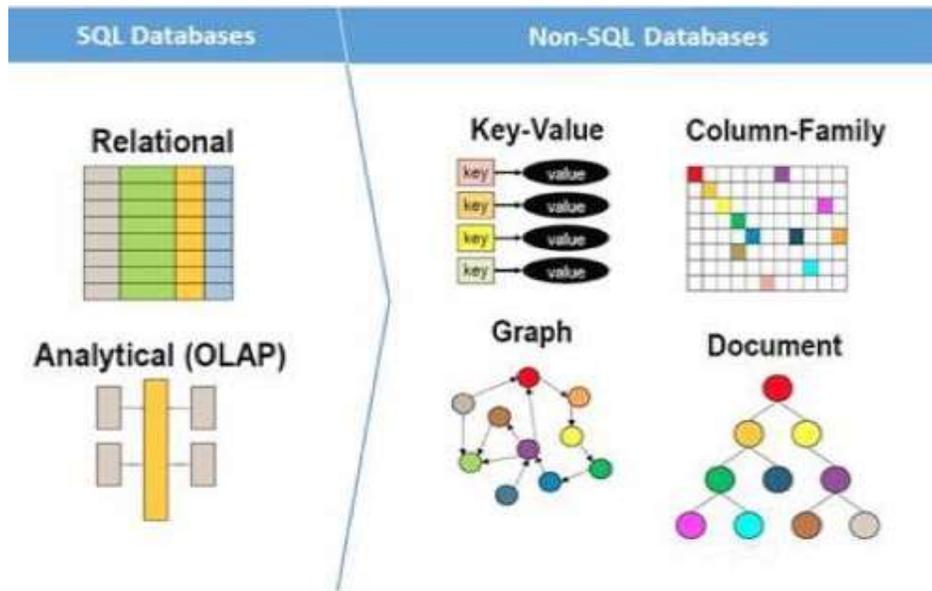
### **2.1.2.3 NoSQL**

Desde su aparición a comienzos de 1960, las bases de datos se han convertido en un soporte para la organización de la información, basadas en el crecimiento y la expansión tecnológica de la

época y a la confiabilidad de los procesos computacionales. Para 1970 se propuso por primera vez el modelo de bases de datos relacionales y las teorías subyacentes, lo que implicó un cambio extremo en el manejo de la información, debido a que los datos comenzaron a ser distribuidos en distintas tablas, con el objetivo de poder consultar la información a través de un lenguaje de consulta estructurado denominado SQL (Structured Query Language), soportado en el álgebra relacional, permitiendo realizar consultas de información de forma declarativa y sin ningún tipo de instrucciones detalladas de programación.

En la actualidad este modelo SQL comenzó a presentar dificultades debido a las grandes cantidades de información que se deben almacenar, la baja velocidad de procesamiento y la imposibilidad de tener la información de manera distribuida, lo que llevó a la aparición de un nuevo modelo que pueda satisfacer las necesidades actuales, este nuevo modelo es conocido como NoSQL, teniendo en cuenta un aspecto particular, y es que, la interfaz de consulta NoSQL no es soportada sobre la interfaz de SQL [44].

Las bases de datos NoSQL (no Only SQL o no relacionales) están diseñadas específicamente para modelos de datos específicos y tienen esquemas flexibles para crear aplicaciones modernas. Las bases de datos NoSQL son ampliamente reconocidas porque son fáciles de implementar, por su funcionalidad y el rendimiento a escala, se utiliza una variedad de modelos de datos para acceder y administrar datos. Estos tipos de bases de datos están optimizados específicamente para aplicaciones que requieren grandes volúmenes de datos, baja latencia y modelos de datos flexibles, lo que se logra mediante la flexibilización de algunas de las restricciones de coherencia de datos de otras bases de datos [42].



*Ilustración 3 Estructura bases de datos SQL y NoSQL<sup>3</sup>*

De las principales características resaltables de una base SQL es que utiliza principalmente un sistema ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad por sus siglas en inglés), donde prima la rigidez forzando en todo momento la consistencia de los datos al final de cada proceso sacrificando la disponibilidad.

Por otra parte, aunque las bases de datos NoSQL no garantiza el sistema ACID, se enfoca en el sistema BASE que prioriza la disponibilidad todo el tiempo, un estado flexible y consistencia eventual, por lo que en un preciso momento puede que no sea consistente, pero a lo largo del tiempo será eventualmente consistente.

En conclusión, las bases de datos NoSQL no se enfocan en unos principios rígidos, sino que se centran en romper con los paradigmas tradicionales y así ganar disponibilidad perdiendo parte de la consistencia, otorgando un estado de los datos más flexible.

<sup>3</sup> <https://zimozi.co/nosql-db-not-only-sql/>

Un teorema que explica la importancia de estos sistemas es el teorema de CAP que dice que en un sistema distribuido con datos compartidos no se puede tener Consistencia, Disponibilidad y tolerancia a particiones. Por lo que debe de optar por dos de las tres características, que mejor se adapte para el sistema distribuido.

- CA: Consistencia y Disponibilidad.
- CP: Consistencia y tolerancia a particiones.
- AP: Disponibilidad y tolerancia a particiones.

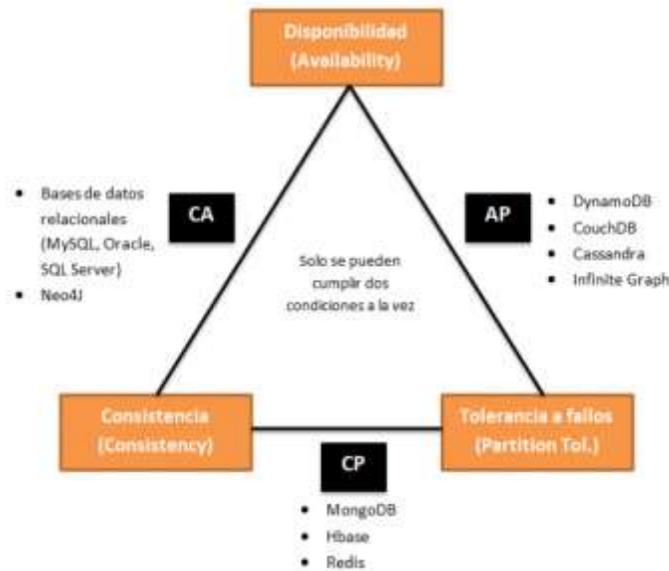


Ilustración 4 Teorema CAP<sup>4</sup>

#### 2.1.2.4 MongoDB

La información antes de ser analizada debe ser almacenada de manera correcta, para este propósito se utilizó MongoDB, considerada actualmente como la base de datos NoSQL más

<sup>4</sup> <https://www.genbeta.com/desarrollo/nosql-clasificacion-de-las-bases-de-datos-segun-el-teorema-cap>

popular. MongoDB es una base de datos orientada a documento, esto quiere decir que el concepto de fila que se maneja usualmente en las bases de datos SQL se reemplaza acá por el concepto de documento, lo que permite almacenar información compleja sin la necesidad de preocuparse por la manera de representar la información en un formato concreto de tablas, lo que sí es obligatorio en las bases de datos SQL.

MongoDB almacena un conjunto de documentos en un espacio llamado colecciones, estos espacios no tienen un esquema prefijado, lo que significa que un documento puede tener unas claves mientras que el siguiente documento puede contener una estructura totalmente diferente.

Una característica más que diferencia a MongoDB de las demás bases de datos es la posibilidad de almacenar grandes volúmenes de información en un entorno escalable, lo que permite que la información se pueda guardar en un clúster de computadores, donde los datos se encuentran repartidos por todo el clúster. En el caso de que la colección siga creciendo, no habría problema con añadir nuevos computadores para aumentar la capacidad de almacenamiento. MongoDB puede ser instalado y ejecutado en una instancia en un servidor cloud como AWS.

MongoDB no presenta problemas de rendimiento en la medida en que la cantidad de información crece, debido a que la velocidad de lectura y búsqueda de datos se realiza en paralelo en todos los computadores donde se encuentren almacenadas las colecciones, lo que brinda la posibilidad de integrar escalabilidad en el tiempo de acceso<sup>5</sup>. [21]

---

<sup>5</sup> <https://www.mongodb.com/>

### 2.1.3 Extract, Transform and Load ETL

En esta etapa se realizan tres subprocesos (extracción, transformación y carga) [28], estos subprocesos se consideran secuenciales, lo que significa que cada uno de ellos suministra la materia prima necesaria para continuar desarrollando el siguiente. El producto final del proceso ETL contiene la base de datos depurada para iniciar el análisis del estudio.

- **Extracción:** este subproceso permite extraer la información contenida en las bases de datos, logrando visualizar la cantidad total de información correspondiente al estudio realizado.
- **Transformación:** en el subproceso de transformación se organiza y unifica la base de datos, se quitan los campos que no se necesitan en la investigación y se realiza imputación de datos si es el caso.
- **Carga:** en este subproceso se extrae la data con la que se va a trabajar en el proceso de investigación, y se carga en los respectivos programas o aplicativos en los que se va a trabajar.

Cada una de las bases resultantes en cada subproceso se guarda en un lago de datos (data lake).

#### 2.1.3.1 Python

Python es un lenguaje de programación que tuvo sus orígenes al inicio de la década de los noventa. El creador fue el holandés e investigador Guido Van Rossum, quien se propuso utilizar un lenguaje de programación distinto a los sugeridos por el medio, con la diferencia de que quería crear su propio lenguaje.

La primera versión de Python fue la 0.9, que a pesar de tener carencias ya comenzaba a marcar diferencia en el año de 1991. Sólo hasta el año 1994 se pudo ver la versión 1.0.

Actualmente Python tiene dos versiones generales, la 2.x y la 3.x, en donde la X determina la secuencia de versiones que se liberan de manera progresiva y secuencial en cada versión general. Entre las versiones generales existen algunas diferencias, pero ha sido pertinente seguir brindando soporte sobre la versión 2.x porque aún existen muchos proyectos exitosos que fueron desarrollados bajo esta versión y no han sido migrados a la versión 3.x.



*Ilustración 5 Logo Python.<sup>6</sup>*

Python es un lenguaje interpretado, lo que significa que necesita un intérprete para leer línea a línea el código escrito y además es un lenguaje de alto nivel, lo que permite que las instrucciones sean más simples y que cumplan cada una con varias funciones al mismo tiempo. Es un lenguaje multiplataforma y algunos sistemas operativos ya traen este lenguaje por defecto [18].

Para el procesamiento de archivos .csv en Python se usa la biblioteca que tiene el mismo nombre antecedido de la función import (`>>> import csv`).

### **2.1.3.2 R y R Studio**

R es un lenguaje de programación con un entorno para análisis estadístico y gráfico, concebido como un proyecto de software libre que tuvo sus inicios en el lenguaje usado a nivel científico

---

<sup>6</sup> <https://www.python.org/>

conocido como S-Plus. Estos dos lenguajes son los más utilizados dentro de la comunidad de investigación informática y estadísticas. R permite cargar diversas bibliotecas o paquetes de cálculo matemático y visualización gráfica distribuidos bajo la licencia GNU GPL y se encuentra disponible para varios sistemas operativos. Además, permite el análisis de grandes cantidades de datos, siendo una herramienta idónea para la minería de datos.

R posee una gran cantidad de herramientas estadísticas para procesar datos de modelos lineales y no lineales, análisis de series temporales, algoritmos de clasificación, agrupamiento y gráficos.

R permite además que los usuarios tengan la libertad de crear sus propias funciones, crear enlaces con bases de datos y con lenguajes de programación como Perl y Python, tanto así, que por su poder de procesamiento numérico puede ser comparado con otras herramientas similares como GNU Octave y MATLAB.



*Ilustración 6 Logo R<sup>7</sup>*

Otros beneficios de R es que posee una interfaz de interacción con Weka (RWeka) para hacer uso de algoritmos especializados en minería de datos.

---

<sup>7</sup> <https://www.r-project.org/>

R Studio es un entorno complementario de R Console y R Commander, que también es software libre y puede ser obtenido desde <https://www.rstudio.com/products/rstudio/download/>. [20]

### **2.1.3.3 Data Lake y Data Warehouse [10], [11]**

Un lago de datos (Data lake) se refiere a un repositorio de almacenamiento masivo y escalable que contiene una gran cantidad de datos sin procesar en su formato nativo ("tal cual") hasta que se necesitan más sistemas de procesamiento (motor) que pueden ingerir datos sin comprometer la estructura de datos. Los lagos de datos generalmente se construyen para manejar grandes volúmenes de datos no estructurados que llegan rápidamente (en contraste con los datos altamente estructurados de los almacenes de datos o data warehouse) de los cuales se derivan más conocimientos. Por lo tanto, los lagos utilizan aplicaciones analíticas dinámicas (no preconstruidas estáticas como en los almacenes de datos). Los datos en el lago se vuelven accesibles tan pronto como se crean [10].

En la industria los Data Lakes son soluciones de gestión de datos híbridos que pueden hacer frente a los retos de big data y que impulsan nuevos niveles de analítica en tiempo real. Con un entorno altamente escalable, soportando volúmenes de datos extremadamente grandes, y acepta datos en su formato nativo a partir de varios orígenes de datos. Como complementos para la data warehouse, proporciona plataforma para machine learning y analítica avanzada en tiempo real en un entorno colaborativo.

Los Data Lakes consolidan los datos en un entorno gobernado y bien administrado que admite tanto el desarrollo de análisis como las cargas de trabajo de producción. Abarca múltiples plataformas de datos tales como relational data warehouses, apache hadoop clusters y dispositivos analíticos, y los gestiona juntos a través de un programa de gobernanza común.

Estas plataformas de datos se pueden distribuir geográficamente. El acceso a las plataformas de datos está restringido a los servicios del Data Lake y los motores que administran los datos. Las aplicaciones y las personas acceden a los datos a través de los servicios del Data Lake.

Data Warehouse es una combinación de tecnologías y componentes para el uso estratégico de datos. Recopila y gestiona datos de diversas fuentes para proporcionar información empresarial significativa. Es el almacenamiento electrónico de una gran cantidad de información diseñada para consultas y análisis en lugar de procesamiento de transacciones. Es un proceso de transformación de datos en información [11].

De las principales diferencias entre data Warehouse y data lake es:

1. Data Lake almacena todos los datos independientemente de la fuente y su estructura, mientras que Data Warehouse almacena los datos en métricas cuantitativas con sus atributos.
2. Data Lake es un repositorio de almacenamiento que almacena enormes datos estructurados, semiestructurados y no estructurados, mientras que Data Warehouse es una combinación de tecnologías y componentes que permite el uso estratégico de datos.
3. Data Lake define el esquema después de que se almacenan los datos, mientras que Data Warehouse define el esquema antes de que se almacenen los datos.
4. Data Lake usa el proceso ELT (Extract Load Transform) mientras que Data Warehouse usa el proceso ETL (Extract Transform Load).
5. Comparando Data Lake vs Warehouse, Data Lake es ideal para aquellos que desean un análisis en profundidad, mientras que Data Warehouse es ideal para usuarios operativos.

6. Los data warehouse consisten en datos extraídos de sistemas transaccionales o se realiza un proceso previo para estructurar los datos y ser almacenados. Los data lake son de fuentes de datos no tradicionales como: registros del servidor web, sensor data (lo que utiliza en su mayoría las IoT), actividad de redes sociales, el texto y las imágenes son ignorados en gran parte. Un Data Lake abarca estos tipos de datos no tradicionales.

#### **2.1.4 Estadística y analítica de datos**

Gran parte de la metodología estadística que se usa en ciencia de datos fue desarrollada e implementada hace más de 50 años, pero solamente hasta la actualidad, en donde se dispone de grandes cantidades y volúmenes de información y de buena potencia de cálculo de los computadores se ha podido poner en práctica.

Un computador personal actual puede manejar volúmenes considerables de información, incluso se puede recurrir a herramientas en la nube a un bajo costo, por tanto, en el campo de la ciencia de datos, bajo la necesidad de suplir el buen funcionamiento de todas las herramientas y plataformas requeridas, se ha visto el surgimiento de distintos roles que juegan un papel diferencial e importante en cada una de las etapas que se necesitan para poder realizar un análisis de datos objetivo [45].

##### **2.1.4.1 Roles**

Los roles en la ciencia de datos van de la mano de la etapa de la analítica de datos en la que se desee incurrir de manera casi secuencial y de la madurez que adquiera la información que se requiere analizar, por tanto, en cada una de estas etapas se pueden hacer visibles distintas categorías de personas o expertos en el análisis de datos, entre ellos se encuentran los científicos

de datos, ingenieros de datos, arquitectos de datos y analistas de datos. El presente trabajo se enfoca en la labor del científico de datos ligada a un análisis descriptivo de los datos.

#### **2.1.4.1.1 Científico de datos**

La profesión del científico de datos es una actividad que hace una década no se encontraba dentro de los radares de las organizaciones como posibles trabajadores a contratar en sus filas, pero debido al crecimiento de la información de cualquier empresa, se hizo necesario contar con personal capacitado en el procesamiento de esta información, con el objetivo de convertirla en conocimiento y que les permita tener una toma de decisiones más acertada.

El científico de datos posee conocimientos en áreas como las matemáticas, estadística, computación y en parte son muy buenos observadores especialmente del comportamiento y la tendencia de los datos.

El científico de datos puede incursionar tanto en pequeñas como en grandes empresas, eso quiere decir que no siempre es necesario que exista una cantidad de información enorme o desbordante, pero para que el científico de datos pueda realizar bien su trabajo necesita de las siguientes habilidades:

- Entender los datos: el científico de datos debe saber que son y que representan los datos que está procesando, de ahí que las técnicas y algoritmos que conocen pueden no ser útiles si no se cuenta con la intuición y la experiencia necesarias para llevar a cabo este primer objetivo, que lo conducirá a comprender cuál es el problema que se desea o se debe resolver. En este punto el científico de datos debe procurar por recolectar los datos que posee y convertirlos en un formato más utilizable

- Comprender el problema a resolver: en esta etapa el científico de datos debe saber que conocimiento se puede extraer de los datos que tiene, es acá donde comienza a jugar un papel importante la estadística, con la que se puede comenzar a determinar cuáles son los patrones de comportamiento de los datos y sus correlaciones. Esta etapa le ayudará al científico de datos a elaborar un modelo que inicialmente le permita conocer bien los datos y así conseguir seleccionar después la información que realmente necesita para finalmente comprobar el ajuste de dicho método.
- Conocer la tecnología disponible: principalmente el científico de datos debe tener un buen conocimiento de la tecnología que está a su alcance y de poderla optimizar para resolver las dificultades que se le presentan, teniendo en cuenta dos variables como son el tiempo y la manera de resolución de los problemas. Entre las herramientas que tiene a su alcance de manera casi gratuita están plataformas como R, Python, Weka y grandes ecosistemas para el manejo de grandes cantidades de información como Hadoop.

En conclusión, ser experto en las 3 habilidades es una labor titánica, debido a esto se han creado equipos de trabajo interdisciplinarios que conllevan a que coexistan otros roles dentro de la ciencia de datos.

#### **2.1.4.1.2 Ingeniero de datos**

El rol del ingeniero de datos es el de convertir en un producto que incluso pueda ser comercializado el trabajo realizado por el científico de datos. En profundidad, el ingeniero de datos es el encargado de la adquisición, almacenamiento, transformación y gestión de datos en una entidad, además asume la configuración de la infraestructura tecnológica necesaria para que un enorme volumen de datos no estructurados recogidos se convierta en un producto que pueda

ser utilizado por otros especialistas de la ciencia de datos como son el científico y el analista de datos.

Los ingenieros de datos diseñan, crean y mantienen la arquitectura de las bases de datos y de los sistemas de procesamiento, de tal manera que, la tarea de explotación, análisis e interpretación de los datos se pueda hacer sin incidencias y de manera segura. En pocas palabras, el ingeniero de datos se centra en el proceso ETL y mantiene toda la infraestructura por donde viajan los datos de manera impecable.

#### **2.1.4.1.3 Arquitecto de datos**

La labor del arquitecto de datos es la más técnica dentro del conjunto de roles de la ciencia de datos y del Big Data, ya que es el encargado de crear la infraestructura tecnológica necesaria para soportar todo el proceso de los datos, incluyendo la recopilación, lectura y salida de estos, por tanto, el arquitecto de datos crea un conjunto de modelos, reglas y estándares que dicen cómo almacenar y organizar los datos para que sean útiles en el proceso de toma de decisiones. Una buena arquitectura de datos le permite al científico de datos trabajar de manera eficiente con datos relevantes y muy confiables.

Anteriormente el trabajo de los arquitectos de datos era realizado por los ingenieros, especialistas en ETL o analistas de datos, lo que generaba que la arquitectura de los datos tuviese enormes brechas de información, causando problemas muy difíciles de solucionar posteriormente.

Otras habilidades de los arquitectos de datos son tener conocimientos en componentes de arquitecturas de datos como Hadoop, competencias en el manejo de bases de datos como modelado y optimización de estas, experiencia con entornos y plataformas cloud tales como

AWS, conocimientos en procesos ETL, comprensión de la metodología SCRUM y experiencia en gestión de proyectos.

#### **2.1.4.1.4 Analista de datos**

El analista de datos tiene como objetivo dentro del ciclo de la información de una empresa, identificar patrones de comportamiento dentro de grandes volúmenes de datos, con el fin de que con estos resultados se pueda llevar a cabo una buena toma de decisiones. Las principales actividades del analista dentro del ciclo de los datos son la posibilidad de extraer, procesar y almacenar datos, además hacer análisis y crear informes que ayudan a interpretar los datos con el fin de establecer modelos predictivos.

En este rol también se incluye la visualización de los datos a través de presentaciones que permitan tener una buena relación entre la información obtenida y las conclusiones esperadas. También se incluyen las habilidades analíticas y aptitudes para la resolución de problemas.

#### **2.1.4.2 Madurez de los datos**

La madurez de los datos o niveles de madurez se mide por el logro de objetivos para garantizar la calidad de la información en términos de valor y beneficios para una organización o estudio de estos. Así entre mayor sea su valor o nivel de madurez mayor es la complejidad que tiene su implementación.



Ilustración 7 Nivel de madurez de los datos valor y complejidad<sup>8</sup>

Estos niveles de madurez generan para una organización o estudio un valor que apoya en la toma de decisiones y pronosticar en un nivel más avanzado, que puede llegar a ocurrir y ser una base para la toma de acción. Puesto que cada nivel resuelve las preguntas de ¿qué ocurrió?, ¿Por qué ocurrió?, ¿Qué ocurrirá? y ¿qué se debe hacer para que pase?



Ilustración 8 Madurez de datos decisión y acción<sup>9</sup>

<sup>8</sup> <https://www.integratetecnologia.es/la-innovacion-necesaria/el-nivel-de-madurez-de-las-organizaciones-en-el-analisis-de-datos/>

<sup>9</sup> <http://www.edcontrol.com/index.php/elementos/elementos-18/item/1-desmitificar-los-conceptos-de-analitica-nube-e-iiot>

#### **2.1.4.2.1 Analítica descriptiva**

El análisis descriptivo se ocupa de estudiar el pasado. Como el nombre lo indica, el análisis descriptivo se usa para describir el histórico en un conjunto de datos. Hay varias formas de describir el pasado:

- Usando estadísticas fáciles de entender, cómo: mínimo, máximo, media, mediana, cuartiles, desviación típica, agrupar datos entre otros.
- Usando gráficos que resuman los datos como histogramas, caja de bigotes, diagramas de torta entre otros.
- Las tablas también son un aspecto fundamental del análisis descriptivo.

En el análisis descriptivo juega un papel importante el científico de datos, a través del conocimiento de los datos y poder entender el problema que se necesita resolver o plantear para una futura toma de decisiones dentro de las organizaciones.

Los hallazgos obtenidos en la analítica descriptiva solamente indican si lo que sucedió está bien o mal, pero aún sin explicar por qué.

#### **2.1.4.2.2 Analítica Diagnóstica**

El análisis diagnóstico permite establecer la causa de lo ocurrido, el porqué de los resultados encontrados en los datos. La analítica diagnóstica utiliza los datos del presente y tiene la facultad de poderlos relacionar con el histórico o los resultados obtenidos en la analítica descriptiva.

La analítica descriptiva y la diagnóstica por lo general nos ayudan a responder preguntas sencillas basadas en los hallazgos encontrados en los datos depurados y que posibilitan seguir

con el proceso predictivo y prescriptivo en el camino de la madurez de la información, considerando la madurez como un proceso secuencial en cualquier organización.

La analítica diagnóstica también permite contrastar la información encontrada con otro tipo de información más específica y detallada que la organización haya adquirido en otro momento.

#### **2.1.4.2.3 Analítica Predictiva**

La analítica predictiva se basa en la información del pasado, en el histórico de los datos, a través de los que se puede predecir que puede llegar a ocurrir en un futuro. En esta etapa juega un papel fundamental el Machine Learning y los algoritmos de predicción existentes actualmente.

En esta etapa los datos deben estar muy bien depurados para que las conclusiones y posibles decisiones que se puedan tomar a futuro en una organización sean lo más acertadas posible.

La analítica predictiva hace uso de los hallazgos del análisis descriptivo y diagnóstico y le permite encontrar grupos y excepciones para predecir tendencias futuras. El pronóstico de todas formas es un valor estimado, por tanto, su precisión depende en gran medida de la calidad de los datos y la estabilidad del sistema implementado.

#### **2.1.4.2.4 Analítica Prescriptiva**

La analítica prescriptiva hace uso de la información causal, la que irá a una herramienta de reconocimiento que es la encargada de indicar que es lo que realmente se debe hacer. La analítica prescriptiva es un soporte fundamental en la toma y automatización de decisiones, suelen involucrar aprendizaje de máquina e inteligencia artificial, además de otro tipo de información, no sólo del histórico de datos, sino también información externa que le permita trabajar de manera más eficiente con los algoritmos utilizados.

La analítica prescriptiva es el escalón más alto en el proceso de los datos, es allí donde los datos están en su madurez más alta. Relacionando la analítica prescriptiva con la analítica predictiva, la analítica prescriptiva ayuda a planear cómo hacer para que las predicciones planteadas en la anterior etapa lleguen a ocurrir o que permita prevenir algún problema detectado a futuro.

### **2.1.4.3 Imputación de datos**

En la realidad es frecuente encontrar archivos con observaciones deliberadamente no diligenciadas, o improbables en los datos suministrados. Por tanto, en estos casos se genera imputación de datos, para no generar fallas en los procesos estadísticos que requieren datos existentes y que no afecten en gran medida la veracidad de los resultados de la investigación. Así se recomienda realizar un preprocesamiento donde se llenen estos campos (si se tiene manera de determinar el dato faltante) o se asocie a un código para diferenciarlos con los otros datos y así poder trabajar con los datos completos. [43]

### **2.1.4.4 Medidas de dispersión (media, mediana y desviación estándar)**

Las medidas de dispersión nos permiten identificar la forma en que se separan o aglomeran los valores de acuerdo con su representación gráfica. Estas medidas describen la manera como los datos tienden a reunirse de acuerdo con la frecuencia con que se encuentran dentro de la información.

- **Promedio o media**
- **La mediana** es el valor de la variable que ocupa la posición central, cuando los datos se disponen en orden de magnitud. Es decir, el 50% de las observaciones tiene valores

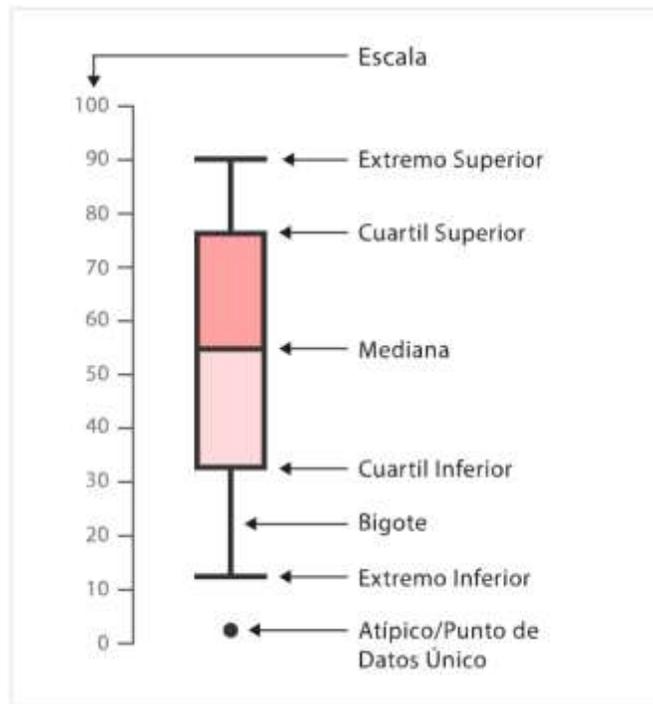
iguales o inferiores a la mediana y el otro 50% tiene valores iguales o superiores a la mediana.

- **La desviación estándar** es un promedio de las desviaciones individuales de cada observación con respecto a la media de una distribución. Así, la desviación estándar mide

el grado de dispersión o variabilidad. 
$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

#### 2.1.4.5 Caja de bigotes (cuartiles, datos atípicos)

Los diagramas de Caja-Bigotes (boxplots o box and whiskers) son una representación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y la simetría. Esta caja se ubica a escala sobre un segmento que tiene como extremos los valores mínimo y máximo de la variable. La gráfica contiene una caja interna donde inicia en el primer cuartil ( $Q_1$ ) hasta el tercer cuartil ( $Q_3$ ) y se divide en el segundo cuartil ( $Q_2$ ), las líneas que sobresalen de la caja se llaman bigotes. Estos bigotes tienen un límite de prolongación, de modo que cualquier dato o caso que no se encuentre dentro de este rango es marcado e identificado individualmente y los datos que se encuentran por fuera de estos extremos se denomina datos atípicos. Los cuartiles son medidas estadísticas de posición que tienen la propiedad de dividir la serie estadística en cuatro grupos de números iguales de términos, el cuartil 2  $Q_2$  es la mediana de los datos, el  $Q_1$  y  $Q_3$  son las medianas de las dos mitades en donde se encuentran.

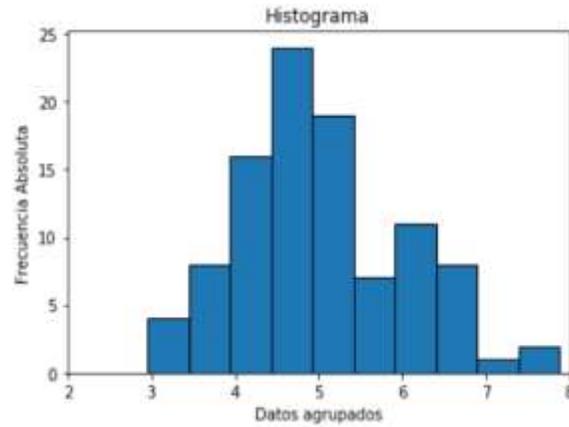


*Ilustración 9 Caja de bigotes<sup>10</sup>*

#### **2.1.4.6 Histogramas**

Un histograma es la representación gráfica en forma de barras, que simboliza la distribución de un conjunto de datos. Sirven para visualizar de manera general la distribución de la población, o de la muestra, respecto a una característica cuantitativa y continua.

<sup>10</sup> [https://datavizcatalogue.com/ES/metodos/diagrama\\_cajas\\_y\\_bigotes.html](https://datavizcatalogue.com/ES/metodos/diagrama_cajas_y_bigotes.html)



*Ilustración 10 Histograma <sup>11</sup>*

#### **2.1.4.7 Análisis multivariado**

El análisis multivariado o la estadística multivariante se refiere a diferentes métodos que estudian y examinan el efecto simultáneo de múltiples variables. Los métodos estadísticos multivariados se utilizan para analizar el comportamiento conjunto de más de una variable aleatoria. Este análisis se propone para medir, explicar y predecir el grado de relación entre la variación.

#### **2.1.4.8 Prueba de hipótesis**

Una prueba de hipótesis es una regla que especifica cuándo se puede aceptar o rechazar una afirmación sobre una población dependiendo de la evidencia proporcionada por una muestra de datos.

Una prueba de hipótesis examina dos hipótesis opuestas sobre una población: la hipótesis nula y la hipótesis alternativa. La hipótesis nula es la afirmación que se está comprobando.

Normalmente la hipótesis nula es una afirmación de "sin efecto" o "sin diferencia". La hipótesis

---

<sup>11</sup> <https://www.superprof.es/apuntes/escolar/matematicas/estadistica/descriptiva/histograma.html>

alternativa es la afirmación que se desea probar para ser capaz de concluir que es verdadera, basándose en la evidencia proporcionada por los datos de la muestra.

Con base en los datos de la muestra, la prueba determina si se puede rechazar la hipótesis nula. Se utiliza el valor  $p$  para tomar esa decisión. Si el valor  $p$  es menor que el nivel de significancia (denotado como  $\alpha$  o alfa), entonces se puede rechazar la hipótesis nula.

## CAPÍTULO III: DESARROLLO DE LA TESIS

### 3.1 Recolectar datos de los resultados de las pruebas Saber 11

En esta primera etapa se determina cuál será la fuente de datos y el tipo de información útil para el desarrollo del modelo.

Por lo tanto, se descargó la información correspondiente a los resultados de las pruebas Saber 11 en Colombia, pertenecientes a los años 2015, 2016, 2017, 2018 y 2019. Este volumen de datos es extraído de las bases de datos del ICFES [5], en formato CSV (valores separados por comas).

El ICFES junto a los puntajes de las pruebas Saber 11 recolecta información adicional a la prueba, tanto del estudiante como de la institución a la que pertenece el estudiante. Una categorización general de dicha información es la siguiente [34][35]:

- Información personal
- Información de contacto
- Información socioeconómica
- Antecedentes escolares
- Expectativas
- Información del colegio
- Datos de citación del examen
- Resultados

Cada uno de estos módulos contienen información única y significativa del estudiante, con la que se pueden realizar distintos análisis no solo de los resultados sino también de la situación del estudiante en su entorno social, familiar y educativo. Es importante mencionar que la data de los periodos 2019-1 y 2019-2 presentaron un cambio con respecto a la data de los periodos entre 2015-1 hasta 2018-2, dichos cambios se ven en los módulos de antecedentes escolares y expectativas, los que se retiran para el año 2019 y se agregan y quitan algunas variables de los otros módulos, por tal motivo es crucial realizar un proceso ETL previo al trabajo con los datos, para así unificar y trabajar de manera conjunta toda la data desde 2015-1 hasta 2019-2.

### **3.2 Realizar el proceso de Extracción, Transformación y Carga (ETL) sobre los datos recolectados de las pruebas Saber 11**

En esta etapa se realizan tres subprocesos (extracción, transformación y carga) [28], estos subprocesos se consideran secuenciales, lo que significa que cada uno de ellos suministra la materia prima necesaria para el siguiente subproceso. Esta fase del proceso es crucial para la investigación puesto que tenemos 10 archivos de datos que contienen información que, aunque para algunos son similares, otros distan en uno y otros como por ejemplo la modificación de la toma de información desde 2019-1.

- **Extracción:** este subproceso permite extraer la información contenida en las bases de datos compartidas por el ICFES, por tanto, se buscaron y descargaron las bases de datos de los años 2015 al 2019, teniendo los datos crudos como los almacena el icfes en .csv, y obteniendo 2 bases de datos por cada año, correspondientes a calendario A y B y nombrados 20xx-1 y 20xx-2 respectivamente, logrando visualizar la cantidad total de información de los estudiantes que presentaron las pruebas Saber 11 durante los años correspondientes al estudio realizado. Las bases de datos del ICFES se encuentran en formato CSV y separando cada prueba en un archivo aparte e independiente de un periodo a otro. No todos los campos o encabezados de las tablas coinciden en todos los archivos obtenidos, por ello se toma el diccionario del ICFES [34][35] para determinar qué campos contiene cada archivo y cómo coinciden entre archivos.

Finalmente, este subproceso arroja 10 archivos de información de los 10 periodos a trabajar (5 años de 2 semestres cada uno).

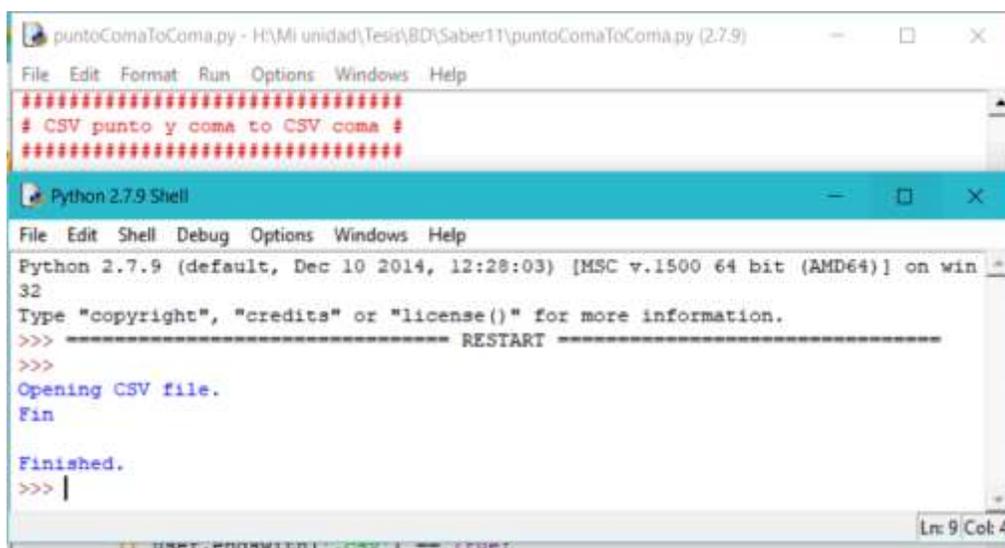
- **Transformación:** en el subproceso de transformación se realizan varios pasos para unificar y organizar toda la información en una sola data de información. Como cada uno de los archivos contiene para el primer periodo un peso de 10 megas y para el segundo periodo 350 megas aproximadamente, al unir todos los archivos se generan una data de 2 gigas aproximadamente de información, sin contemplar que al unirlas se puede incrementar considerablemente el tamaño al combinar los datos, por tal motivo no es posible utilizar métodos comunes como Microsoft Excel para trabajar estos datos, y si se agrega que la estructura de estos archivos no es la misma, no es viable usar los modelos SQL, que para el caso generarían mayores inconvenientes en el manejo de toda la data. Por tal motivo se optó por dar un manejo mediante bases de datos NoSQL, la cual no requiere que los datos tengan la misma estructura y aporta una gran flexibilidad en el manejo de una data con estas condiciones.

Definiendo el proceso de manejo de los datos para unificar y ordenarlos se siguieron los siguientes pasos:

- **Paso 1:** Antes de pasar la data a un procesado de datos NoSQL de debe asegurar que el archivo plano este en un archivo .csv y cada columna este separada por coma, adicionalmente no pueden contener coma dentro de los campos o por lo menos estas comas deben de estar señaladas como delimitador para que el procesador no genere conflicto en su procesamiento, por tal motivo se creó un programa en python el cual:
  - Convertía las comas (,) en guion medio (-), estas comas corresponden a aquellas que estaban contenidas dentro de la información suministrada, no

corresponden a delimitador de elementos y tampoco tiene formato para señalar la exclusión como delimitador.

- Como los delimitadores de estos archivos se encontraban con (;), se reemplazaron con comas (,) con la función de delimitadores, para que quedaran con un formato que se pueda utilizar en MongoDB y el respectivo cargue de archivos .csv.



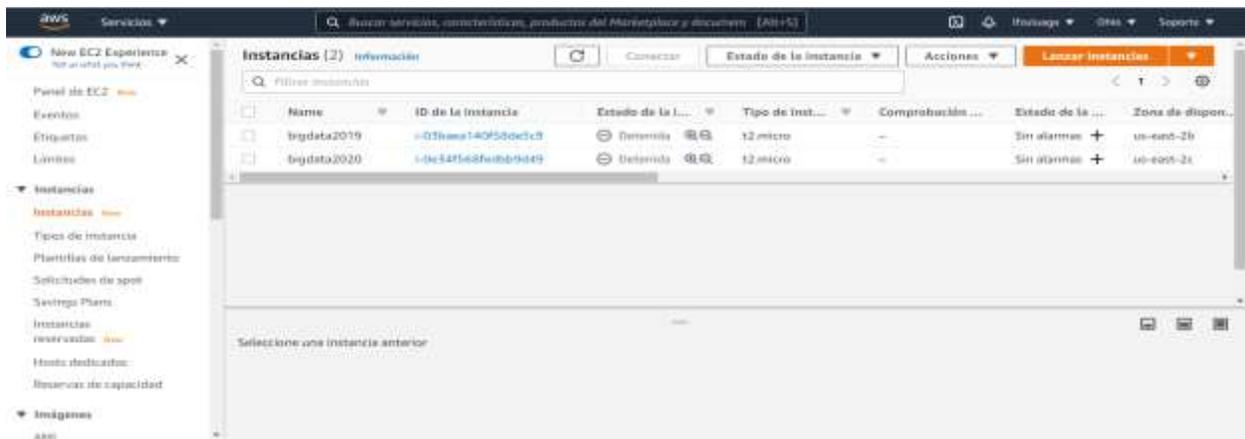
```
puntoComaToComa.py - H:\Mi unidad\Tesis\BD\Saber11\puntoComaToComa.py (2.7.9)
File Edit Format Run Options Windows Help
#####
# CSV punto y coma to CSV coma #
#####

Python 2.7.9 Shell
File Edit Shell Debug Options Windows Help
Python 2.7.9 (default, Dec 10 2014, 12:28:03) [MSC v.1500 64 bit (AMD64)] on win
32
Type "copyright", "credits" or "license()" for more information.
>>> ----- RESTART -----
>>>
Opening CSV file.
Fin

Finished.
>>> |
```

*Ilustración 11 Proceso de transformación de separadores de datos .csv*

- **Paso 2:** Se utilizó un servidor cloud AWS, el que permite la unificación de dichas tablas en una sola, a través de MongoDB, aprovechando los recursos de software y hardware que este tipo de plataformas en línea posee para el procesamiento de información de gran tamaño y así se genera en una base de datos unificada la cual posee un total de 4.294 GB de información.



*Ilustración 12 Panel de Control Instancias AWS*

```

ubuntu@ip-172-31-36-155 ~$
May 22 14:39:31 ip-172-31-36-155 systemd[1]: Started An object/document-oriented database.
ubuntu@ip-172-31-36-155:~$ mongo
MongoDB shell version v3.6.3
connecting to: mongodb://127.0.0.1:27017
MongoDB server version: 3.6.3
Server has startup warnings:
2021-05-22T14:39:32.750+0000 I STORAGE [initandlisten] ** WARNING: Using the XFS filesystem is strongly recommended with
the WiredTiger storage engine
2021-05-22T14:39:32.750+0000 I STORAGE [initandlisten] **          See http://dochub.mongodb.org/core/prodnotes-filesystem
2021-05-22T14:39:37.161+0000 I CONTROL [initandlisten] ** WARNING: Access control is not enabled for the database.
2021-05-22T14:39:37.161+0000 I CONTROL [initandlisten] **          Read and write access to data and configuration is unrestrict
2021-05-22T14:39:37.161+0000 I CONTROL [initandlisten]
> show dbs
admin          0.0000GB
config        0.0000GB
ejemplo       0.0000GB
icfes         4.294GB
local         0.0000GB
>

```

*Ilustración 13 Información de db Icfes en Mongo*

- **Paso 3:** Se realiza un filtro de los campos exclusivamente necesarios para la investigación y se exporta.

```

ubuntu@ip-172-31-36-155: ~
te access to data and configuration is unrestricted.
2021-05-22T14:39:37.161+0000 I CONTROL [initandlisten]
> quit
function quit() {
  [native code]
}
> exit
bye
ubuntu@ip-172-31-36-155:~$ mongoexport --db icfes --collection saber --type=csv
--fields COLE_DEPTO_UBICACION,COLE_MCPPIO_UBICACION,COLE_NATURALEZA,COLE_BILINGUE
,DESEMP_INGLES,PUNT_LECTURA_CRITICA,PUNT_MATEMATICAS,PUNT_C_NATURALES,PUNT_SOCIA
LES_CIUDADANAS,PUNT_INGLES,PUNT_GLOBAL,ESTU_ESTUDIANTE --out /home/ubuntu/saber1
l_d.csv

```

Ilustración 14 Exportar de db Icfes con filtros de Mongo

- **Paso 4:** Se realizó imputación de datos, puesto que se encontraron 40 registros sin información (vacíos) en la variable puntaje de inglés (PUNT\_INGLES). Se calculó el valor de estos registros faltantes realizando una regla de 3 simple, utilizando la siguiente fórmula para el cálculo del puntaje global y despejando el puntaje de

$$\text{Inglés} = Pg \frac{13}{5} - LC * 3 - Mat * 3 - SyC * 3 - CN * 3$$

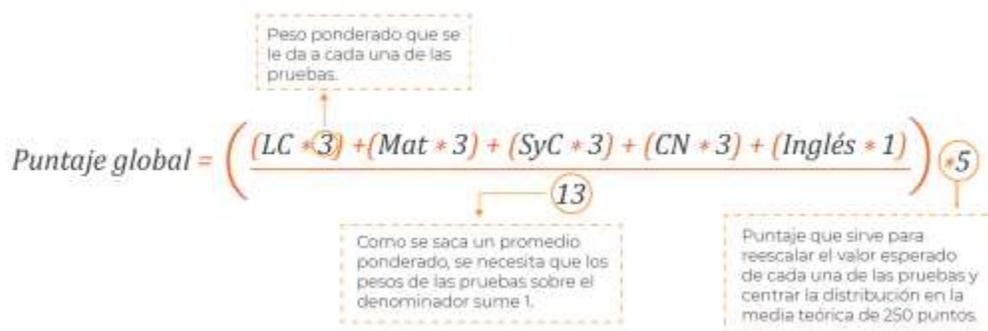


Ilustración 15 Cálculo del Puntaje Global<sup>12</sup>

<sup>12</sup> <https://www.icfes.gov.co/documents/20143/1711757/Informe+nacional+de+resultados+Saber+11-2019.pdf/01cca382-1f24-aefd-a3ef-0d04d2e6108d?version=1.0&t=1608776793757>

- **Carga:** En este subproceso se realizó el cargue del archivo .csv al Rstudio para iniciar como tal el análisis de la información.

### **3.3 Implementar un lago de datos para gestionar los datos crudos pre procesados y procesados**

Después de cargar las bases de datos en el servidor, discriminadas por años, se procede a realizar la unificación de estas en un solo dataset. Esta actividad se realiza utilizando MongoDB y Docker, obteniendo un solo archivo unificado con 2.783.460 registros con 86 columnas correspondientes a los estudiantes que presentaron las pruebas Saber 11 durante los años 2015, 2016, 2017, 2018 y 2019 en Colombia. Durante todo el proceso ETL, se almacena cada uno de los archivos resultantes o intermedio, para realizar una trazabilidad de la información, y tener a la mano información en cada uno de los estados de la data.

### 3.4 Identificar las variables de valor sobre los datos a procesar

Durante el proceso de transformación se generó el proceso de filtrado de campos, conservando todos los datos que responden a la pregunta formulada en el proyecto y definidos como datos de valor o datos importantes para la investigación, lo que nos permitió establecer límites alcanzables para la etapa de análisis del proyecto.

Seguimos conservando la base de datos unificada con 2.783.460 registros y 12 columnas.

- COLE\_DEPTO\_UBICACION
- COLE\_MCPIO\_UBICACION
- COLE\_NATURALEZA
- COLE\_BILINGUE
- DESEMP\_INGLES
- PUNT\_LECTURA\_CRITICA
- PUNT\_MATEMATICAS
- PUNT\_C\_NATURALES
- PUNT\_SOCIALES\_CIUDADANAS
- PUNT\_INGLES
- PUNT\_GLOBAL
- ESTU\_ESTUDIANTE

Las variables se seleccionaron por las siguientes razones:

- Naturaleza del colegio (COLE\_NATURALEZA): este campo tiene dos valores posibles, son NO OFICIAL (privado) y OFICIAL (público).

- **Carácter del estudiante (COLE\_ESTUDIANTE):** este campo tiene tres posibles valores, ellos son ESTUDIANTE, VALIDANTE e INDIVIDUAL, para el proceso de análisis solo se tomarán los registros que sean ESTUDIANTES.
- **Carácter bilingüe (COLE\_BILINGUE):** este campo tiene dos valores posibles, ellos son N (No es bilingüe) y S (es bilingüe).
- **Puntajes por áreas:** existen 5 áreas evaluadas en las pruebas Saber 11, estas son Lectura Crítica (PUNT\_LECTURA\_CRITICA), Matemáticas (PUNT\_MATEMATICAS), Ciencias Naturales (PUNT\_C\_NATURALES), Ciencias Sociales y Competencias Ciudadanas (PUNT\_SOCIALES\_CIUADANAS) e inglés (PUNT\_INGLES). Cada una de estas variables tiene un rango de puntuación entre 0 y 100 puntos.
- **Puntaje global o total (PUNT\_GLOBAL):** esta variable varía entre 0 y 500 puntos, además está categorizada en unos criterios de desempeño cuyos rangos son los siguientes.

<b>Nivel de desempeño global</b>	<b>Rango</b>
Bajo	0-174
Medio	175-249
Medio Alto	250-324
Alto	325-500

*Tabla 2 Rango niveles de desempeño Puntaje Global Pruebas Saber 11*

- Desempeño en el área de inglés (DESEMP\_INGLES): los posibles valores de esta variable son A-, A1, A2, B1 y B+ [38].

<b>Desempeño en inglés</b>	<b>Rango</b>
A-	0-47
A1	48-57
A2	58-67
B1	68-78
B+	79-100

*Tabla 3 Rango niveles de desempeño Puntaje Inglés Pruebas Saber 11*

### 3.5 Buscar patrones de incidencia y valores estadísticos sobre las variables identificadas

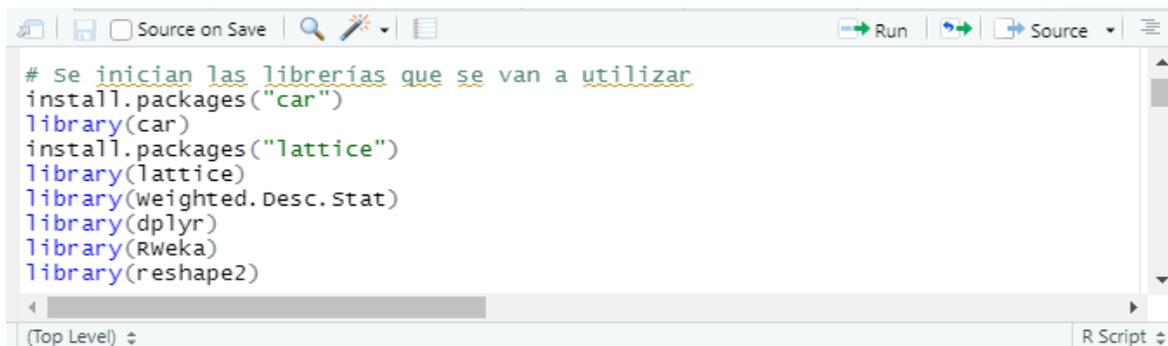
Para iniciar el análisis se toman las variables de valor COLE\_NATURALEZA y

COLE\_BILINGUE como variables para separar las poblaciones.

COLE_BILINGUE	COLE_NATURALEZA	Número de Estudiantes
S	NO OFICIAL	25.085
S	OFICIAL	21.048
N	NO OFICIAL	725.460
N	OFICIAL	2.011.867
Total		2.783.460

*Tabla 4 Número de estudiantes por población privado - público bilingüe*

Las librerías usadas en RStudio para esta etapa de procesamiento de los datos son las siguientes:



```
# se inician las librerías que se van a utilizar
install.packages("car")
library(car)
install.packages("lattice")
library(lattice)
library(weighted.Desc.Stat)
library(dplyr)
library(Rweka)
library(reshape2)
```

*Ilustración 16 Librerías usadas en Rstudio*

A partir de estas dos variables construimos tablas que reflejan la relación que existe entre las poblaciones objeto de estudio y los diferentes desempeños en cada una de las pruebas, así mismo el conglomerado general o puntaje global.

Lectura Crítica	0 -10%	10 - 25%	25 - 50%	50 - 75%	75 - 90%
Público - Total	0-39	40-44	45-51	52-58	59-64 Max 100
Privado - Total	0-41	42-47	42-56	57-63	64-69 Max 100
Público - Bilingüe	0-37	38-42	43-48	49-55	56-61 Max 100
Privado - Bilingüe	26-52	53-59	60-66	67-71	72-76 Max 100
Público - no bilingüe	0-39	40-44	45-51	52-58	59-64 Max 100
Privado - no bilingüe	0-41	42-47	48-55	56-63	64-69 Max 100

*Tabla 5 Agrupación de puntuaciones Lectura crítica*

Matemáticas	0 -10%	10 - 25%	25 - 50%	50 - 75%	75 - 90%
Público - Total	0-35	36-41	42-49	50-57	58-64 Max 100
Privado - Total	0-36	37-44	45-54	55-64	65-72 Max 100
Público - Bilingüe	7-32	33-38	39-46	47-54	55-61 Max 100
Privado - Bilingüe	22-51	52-61	62-69	70-77	78-84 Max 100
Público - no bilingüe	0-35	36-41	42-49	50-57	58-64 Max 100
Privado - no bilingüe	0-36	37-44	45-54	55-63	64-71 Max 100

*Tabla 6 Agrupación de puntuaciones Matemáticas*

Ciencias Naturales	0 -10%	10 - 25%	25 - 50%	50 - 75%	75 - 90%
Público - Total	0-37	38-42	43-49	50-56	57-62 Max 100
Privado - Total	0-38	39-45	46-54	55-62	63-69 Max 100
Público - Bilingüe	17-35	36-40	41-47	48-53	54-60 Max 100
Privado - Bilingüe	19-51	52-60	61-67	68-73	74-79 Max 100
Público - no bilingüe	0-37	38-42	43-49	50-56	57-62 Max 100
Privado - no bilingüe	0-38	39-45	46-53	54-62	63-68 Max 100

*Tabla 7 Agrupación de puntuaciones Ciencias Naturales*

Ciencias Sociales	0 - 10%	10 - 25%	25 - 50%	50 - 75%	75 - 90%
Público - Total	0-34	35-40	41-48	49-56	57-62 Max 100
Privado - Total	0-36	37-43	44-53	54-62	63-70 Max 100
Público - Bilingüe	3-32	33-37	38-44	45-52	53-59 Max 92
Privado - Bilingüe	20-49	50-59	60-67	68-73	74-79 Max 100
Público - no bilingüe	0-34	35-40	41-48	49-56	57-62 Max 100
Privado - no bilingüe	0-36	37-43	44-53	54-62	63-69 Max 100

*Tabla 8 Agrupación de puntuaciones Ciencias Sociales*

Inglés	0 - 10%	10 - 25%	25 - 50%	50 - 75%	75 - 90%
Público - Total	0-36	37-41	42-48	49-54	55-62 Max 100
Privado - Total	0-38	39-45	46-55	56-67	68-78 Max 100
Público - Bilingüe	0-35	36-40	41-46	47-52	53-59 Max 100
Privado - Bilingüe	0-56	57-74	75-81	82-87	88-95 Max 100
Público - no bilingüe	0-36	37-41	42-48	49-54	55-62 Max 100
Privado - no bilingüe	0-38	39-45	46-54	55-66	67-76 Max 100

*Tabla 9 Agrupación de puntuaciones inglés*

Global	0 -10%	10 - 25%	25 - 50%	50 - 75%	75 - 90%
Público - Total	0-188	189-213	214-245	246-278	279-308 Max 478
Privado - Total	0-195	196-227	228-271	272-314	315-348 Max 494
Público - Bilingüe	84-180	181-201	202-230	231-263	264-294 Max 456
Privado - Bilingüe	140-263	264-310	311-344	345-371	372-395 Max 494
Público - no bilingüe	0-189	190-213	214-245	246-278	279-308 Max 478
Privado - no bilingüe	0-194	195-225	226-268	269-311	312-344 Max 492

*Tabla 10 Agrupación de puntuaciones Global*

Revisando el comportamiento de las diferentes áreas, se identifica que tienden hacia una proporcionalidad similar entre sí, por tal motivo se opta únicamente por tomar la puntuación global que conglomerará el total de las puntuaciones.

Como dato relevante, se encontró que, en todas las categorías de las puntuaciones establecidas en las tablas anteriores, en todos los casos las puntuaciones de los colegios públicos y privados bilingües inician en un valor superior a cero, excepto en las puntuaciones de inglés.

Continuamos realizando el estudio con caja de bigotes, para identificar tanto de manera visual como numérica como es el comportamiento de la puntuación global entre las 4 poblaciones.

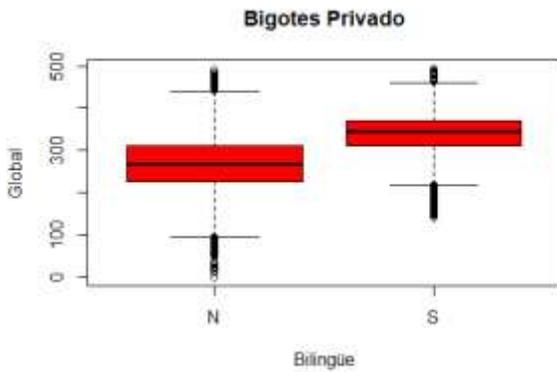


Ilustración 17 Bigotes privado puntaje global

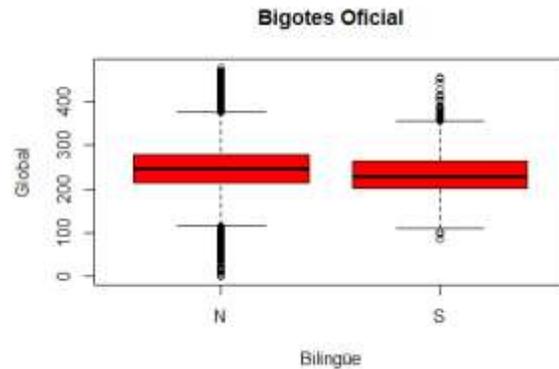


Ilustración 18 Bigotes público puntaje global

Se encuentra que hay una sustancial diferencia en los puntajes de privado bilingüe con respecto a las otras poblaciones. Igualmente se identifica que las poblaciones público bilingüe y público no bilingüe se comportan de maneras muy similares, por tal motivo, se tomó como una sola población a todos los estudiantes de colegios públicos (oficiales).

Considerando lo anterior, se obtiene una tabla comparativa de las diferentes poblaciones encontradas para el análisis de estas de la siguiente manera:

Puntaje Global	0 -10%	10 - 25%	25 - 50%	50 - 75%	75 - 90%
Público - Total	0-188	189-213	214-245	246-278	279-308 Max 478
Privado - Bilingüe	140-263	264-310	311-344	345-371	372-395 Max 494
Privado - no Bilingüe	0-194	195-225	226-268	269-311	312-344 Max 492

Tabla 11 Porcentaje de puntuaciones Global por Colegios

De esta forma, encontramos de manera más clara las diferencias entre los puntajes de los estudiantes pertenecientes a colegios oficiales bilingües y no bilingües (Público - Total), lo que

muestra que el promedio de los puntajes de los colegios oficiales bilingües se encuentra por debajo del promedio general de los colegios oficiales.

Se han considerado también los datos atípicos (outliers) iniciando con los estudiantes de colegios oficiales, donde podemos decir que la cantidad de estudiantes encontrados en estos sectores de la población son pocos en proporción comparado con la cantidad total de la población. Los siguientes son los resultados:

Límite	Puntaje	Porcentaje	Total de estudiantes
Inferior	$213 - (278-213) * 1.5 = 97.5$	0.0078%	159/2.032.915
Superior	$278 + (278-213) * 1.5 = 375.5$	0.25%	5.099/2.032.915

*Tabla 12 Promedio datos atípicos de colegios públicos*

Límite inferior =  $213 - (278-213) * 1.5 = 97.5$

$p = 7.821281e-05 \Rightarrow 0.0078\%$

159

Límite superior =  $278 + (278-213) * 1.5 = 375.5$

$p = 0.002508221 \Rightarrow 0.25\%$

5.099

Se han considerado también los datos atípicos (outliers) de los estudiantes de colegios privados bilingües, donde encontramos que la proporción de estudiantes privados bilingües atípicos es considerablemente más alta que la encontrada en los estudiantes de colegios públicos, debido a

que los límites inferior y superior difieren en cada caso entre 90 y 120 puntos. Los siguientes son los resultados:

Límite	Puntaje	Porcentaje	Total de estudiantes
Inferior	$310 - (371-310) * 1.5 = 218.5$	3.55%	891
Superior	$371 + (371-310) * 1.5 = 462.5$	0.155%	39

*Tabla 13 Promedio datos atípicos de colegios privados bilingües*

Promedio datos atípicos Privado Bilingüe

Límite inferior =  $310 - (371-310) * 1.5 = 218.5$

$p = 0.03551923 \Rightarrow 3.55\%$

891

Límite superior =  $371 + (371-310) * 1.5 = 462.5$

$p = 0.001554714 \Rightarrow 0.155\%$

39

De igual manera se realiza el análisis de los datos atípicos de los estudiantes de colegios privados no bilingües donde se encuentra que la proporción es despreciable, siguiendo el patrón de los estudiantes de colegios públicos con respecto al límite inferior y conservando estrecha similitud con los colegios privados bilingües en el ítem límite superior, encontrando los siguientes resultados:

Límite	Puntaje	Porcentaje	Total de estudiantes
Inferior	$225 - (311-225) * 1.5 = 96$	0.016%	119
Superior	$311 + (311-225) * 1.5 = 440$	0.069%	503

*Tabla 14 Promedio datos atípicos de colegios privados no bilingües*

Promedio datos atípicos Privado no Bilingüe

Límite inferior =  $225 - (311-225) * 1.5 = 96$

$p = 0.0001640339 \Rightarrow 0.016\%$

119

Límite superior =  $311 + (311-225) * 1.5 = 440$

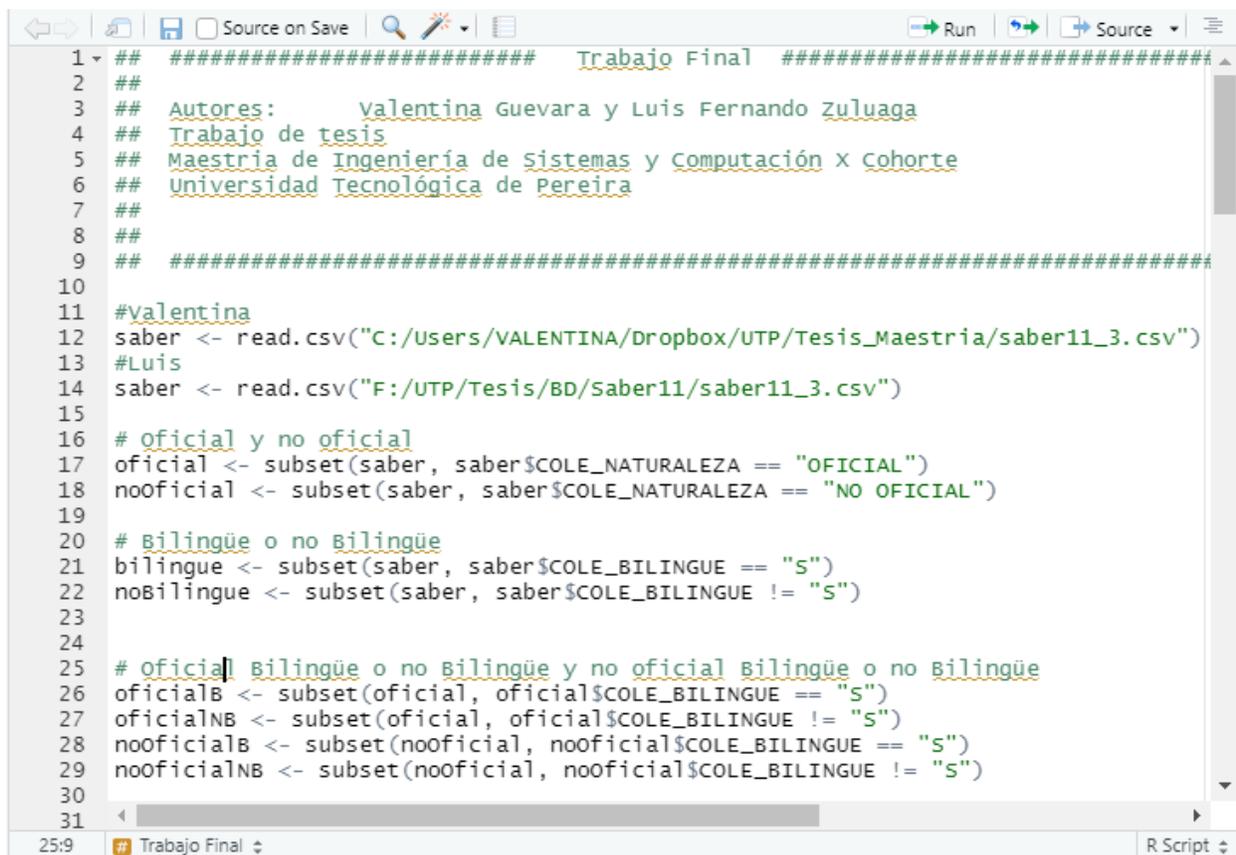
$p = 0.0006933532 \Rightarrow 0.069\%$

503

### 3.6 Cargar datos en modelo de análisis multivariado

Para el análisis multivariado de datos se utilizó el lenguaje R a través de R Studio, en el que se realizó el cargue de la base de datos. Teniendo en cuenta el avance del estudio y las variables más significativas, se realizó un filtro teniendo en cuenta la naturaleza del colegio y si es o no bilingüe. Con estos filtros se implementaron las gráficas denominadas cajas de bigotes, para tener un comparativo entre las distintas subcategorías consideradas a partir de las variables más significativas.

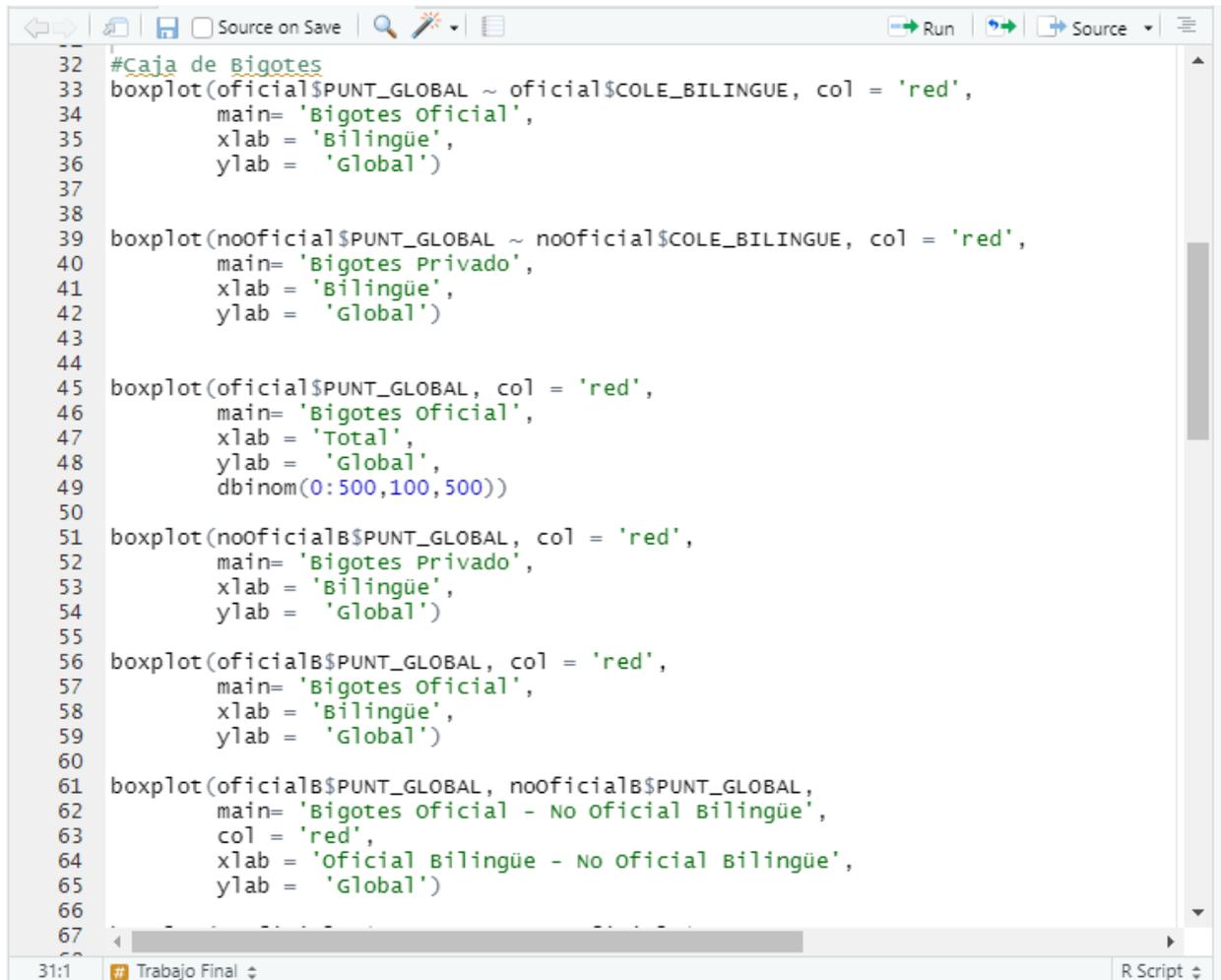
Estos filtros se realizaron de la siguiente manera:



```
1  ## ##### Trabajo Final #####
2  ##
3  ## Autores:      Valentina Guevara y Luis Fernando Zuluaga
4  ## Trabajo de tesis
5  ## Maestría de Ingeniería de Sistemas y Computación x Cohorte
6  ## Universidad Tecnológica de Pereira
7  ##
8  ##
9  ## #####
10
11 #valentina
12 saber <- read.csv("C:/Users/VALENTINA/Dropbox/UTP/Tesis_Maestria/saber11_3.csv")
13 #Luis
14 saber <- read.csv("F:/UTP/Tesis/BD/Saber11/saber11_3.csv")
15
16 # oficial y no oficial
17 oficial <- subset(saber, saber$COLE_NATURALEZA == "OFICIAL")
18 nooficial <- subset(saber, saber$COLE_NATURALEZA == "NO OFICIAL")
19
20 # Bilingüe o no Bilingüe
21 bilingue <- subset(saber, saber$COLE_BILINGUE == "S")
22 nobilingue <- subset(saber, saber$COLE_BILINGUE != "S")
23
24
25 # oficial Bilingüe o no Bilingüe y no oficial Bilingüe o no Bilingüe
26 oficialB <- subset(oficial, oficial$COLE_BILINGUE == "S")
27 oficialNB <- subset(oficial, oficial$COLE_BILINGUE != "S")
28 nooficialB <- subset(nooficial, nooficial$COLE_BILINGUE == "S")
29 nooficialNB <- subset(nooficial, nooficial$COLE_BILINGUE != "S")
30
31
```

*Ilustración 18 Carga y filtro de datos en R Studio*

Las gráficas o cajas de bigotes obtenidas en RStudio se originaron a partir de la separación de las poblaciones anteriormente mencionadas, utilizando la función “boxplot” y las propiedades del entorno gráfico del software.

The image shows a screenshot of the RStudio interface. The main window displays R code for creating boxplots. The code is as follows:

```
32 #Caja de Bigotes
33 boxplot(oficial$PUNT_GLOBAL ~ oficial$COLE_BILINGUE, col = 'red',
34         main= 'Bigotes Oficial',
35         xlab = 'Bilingüe',
36         ylab = 'Global')
37
38
39 boxplot(nooficial$PUNT_GLOBAL ~ nooficial$COLE_BILINGUE, col = 'red',
40         main= 'Bigotes Privado',
41         xlab = 'Bilingüe',
42         ylab = 'Global')
43
44
45 boxplot(oficial$PUNT_GLOBAL, col = 'red',
46         main= 'Bigotes Oficial',
47         xlab = 'Total',
48         ylab = 'Global',
49         dbinom(0:500,100,500))
50
51 boxplot(nooficialB$PUNT_GLOBAL, col = 'red',
52         main= 'Bigotes Privado',
53         xlab = 'Bilingüe',
54         ylab = 'Global')
55
56 boxplot(oficialB$PUNT_GLOBAL, col = 'red',
57         main= 'Bigotes Oficial',
58         xlab = 'Bilingüe',
59         ylab = 'Global')
60
61 boxplot(oficialB$PUNT_GLOBAL, nooficialB$PUNT_GLOBAL,
62         main= 'Bigotes Oficial - No Oficial Bilingüe',
63         col = 'red',
64         xlab = 'Oficial Bilingüe - No oficial Bilingüe',
65         ylab = 'Global')
66
67
```

The RStudio interface includes a toolbar at the top with icons for navigation, saving, and running code. The status bar at the bottom shows the file name 'Trabajo Final' and the current script type 'R Script'.

*Ilustración 19 Caja de Bigotes en R Studio*

Con los resultados obtenidos en el numeral 3.5 de este documento, se encontró que el 90% de los estudiantes de colegios públicos en Colombia no supera en promedio un puntaje global de 308 puntos en las pruebas Saber 11 y el 10% de esta población está por debajo de 188 puntos.

De igual manera, se realizó un comparativo entre los puntajes globales obtenidos por los estudiantes de los colegios privados (No oficiales), diferenciando entre bilingües y no bilingües, en el que se logra establecer los puntajes globales máximos que logran obtener estos estudiantes. Las diferencias encontradas entre los puntajes de los estudiantes pertenecientes a colegios privados bilingües y no bilingües muestra que el 90% de los estudiantes de colegios privados bilingües llegan a un máximo de 395 puntos en comparación a los privados no bilingües que llegan en las pruebas Saber 11 a 344 puntos y el 10% de los privados bilingües está por debajo de 263 puntos, pero los colegios privados no bilingües no superan los 194 puntos, mostrando en el límite inferior la gran diferencia existente entre los bilingües y no bilingües de los colegios privados.

Puntaje Global	0 -10%	75 - 90%
Público - Total	0-188	279-308
Privado - Bilingüe	0-263	372-395
Privado - no bilingüe	0-194	312-344

*Tabla 15 Porcentaje comparativo de puntuaciones Global por Colegios*

### 3.7 Crear el modelo descriptivo utilizando analítica de datos sobre las variables de valor

El análisis de datos se realizó sobre la totalidad de la población de estudiantes que presentaron las pruebas Saber 11 entre los años 2015 al 2019, con un preprocesamiento inicial y habiendo seleccionado unas variables de valor que brindaron información relevante al estudio planteado.

La base de datos de los estudiantes que presentaron las pruebas Saber 11 durante estos 5 años es de 2.783.460 registros y 12 columnas. Para el estudio, se comenzó por identificar la cantidad de registros existentes en cada variable de valor y posteriormente se hizo un análisis multivariado, para reconocer patrones de incidencia entre ellas y así llegar a conclusiones concretas dentro del estudio. Los datos son los siguientes:

<b>COLE_NATURALEZA</b>	<b>Estudiantes</b>
NO OFICIAL	750.545
OFICIAL	2.032.915
Total	2.783.460

*Tabla 16 Naturaleza del Colegio*

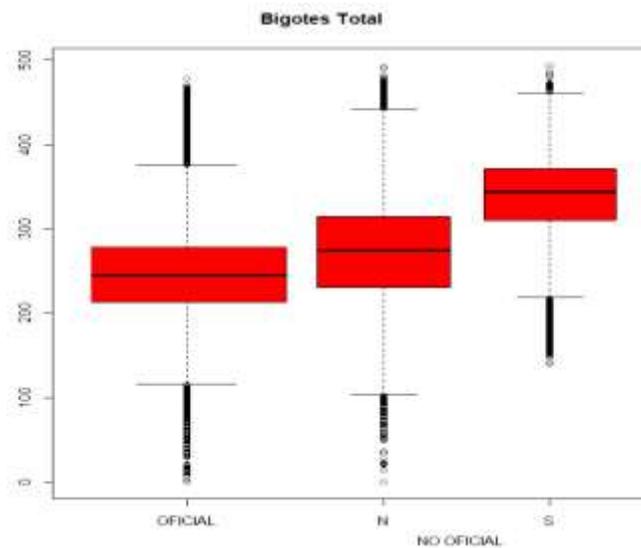
<b>COLE_BILINGUE</b>	<b>Estudiantes</b>
S	46.133
N	2.737.327
Total	2.783.460

*Tabla 17 Colegios Bilingües*

COLE_BILINGUE	COLE_NATURALEZA	Estudiantes
S	NO OFICIAL	25.085
S	OFICIAL	21.048
N	NO OFICIAL	725.460
N	OFICIAL	2.011.867
Total		2.783.460

*Tabla 18 Colegios Bilingües Vs Naturaleza del Colegio*

Seguidamente se realiza el análisis del comportamiento del resultado global con respecto a las 3 poblaciones estudiadas, generando una caja de bigotes comparativa, la que se muestra a continuación:

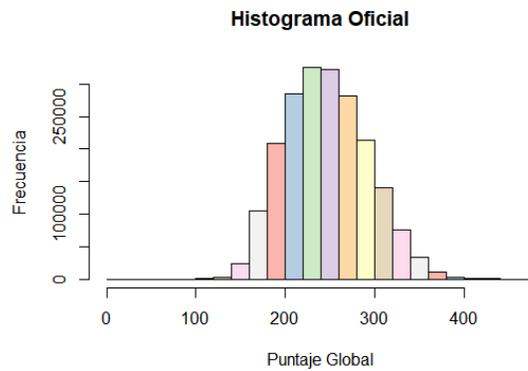


*Ilustración 20 Caja de Bigotes Puntaje Global Oficial vs No Oficial*

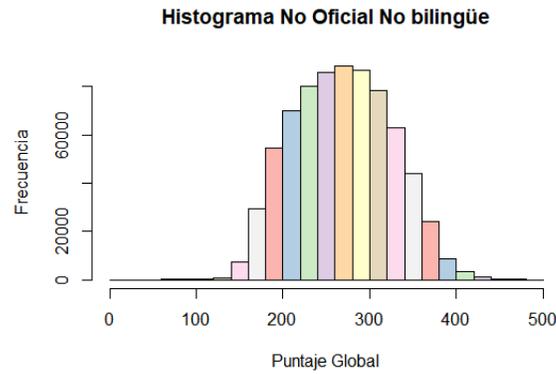
En el anterior gráfico identificamos claramente que son 3 las poblaciones que se comportan de manera distinta entre estas, y según análisis anteriores, los datos atípicos de los estudiantes de colegios públicos (oficiales) y los privados no bilingües no son significativamente distintos.

Adicionalmente se muestra que la población de colegios privados bilingües tiende a tener un mejor resultado en las pruebas globales con relación a las otras dos poblaciones también comparadas.

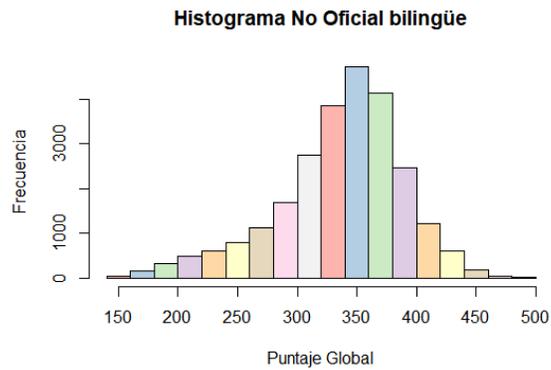
Continuando con el análisis de los datos realizamos los histogramas de las poblaciones, para identificar cómo es el comportamiento de estas e identificar si tienen un comportamiento normal, las gráficas son las siguientes:



*Ilustración 21 Histogramas Puntaje Global publico*



*Ilustración 22 Histogramas Puntaje Global Privado no bilingüe*



*Ilustración 23 Histogramas Puntaje Global Privado Bilingüe*

En las gráficas se observa una distribución normal de los datos a nivel general, donde los histogramas de colegios públicos y privados no bilingües tienden a comportarse de una manera más simétrica comparado con el histograma de colegios privados bilingües, que se encuentra sesgado hacia la derecha.

Identificando los niveles de desempeño global de los resultados, se generó una tabla de rangos donde se establecen los niveles bajo, medio, medio alto y alto de la siguiente manera:

Desempeño	Bajo		Medio		Medio Alto		Alto	
Rango porcentual (%)	0	35%	35%	50%	50%	65%	65%	100%
Rango por puntos	0	174	175	249	250	324	325	500

*Tabla 19 Niveles de Desempeño Puntaje Global Pruebas Saber 11*

Estos niveles nos permiten agrupar el desempeño de los estudiantes y realizar una clasificación cualitativa de estos resultados para un análisis descriptivo y diagnóstico de los datos.

### 3.8 Validar los patrones hallados mediante pruebas estadísticas

Posteriormente identificamos qué porcentaje de estudiantes se encuentra en los diferentes niveles de desempeño establecidos.

Nivel	Privado Bilingüe	Privado No Bilingüe	Público
Bajo (0-174)	0,50%	3,56%	4,30%
Medio (175-249)	7,02%	35,03%	49,57%
Medio Alto (250-324)	26,51%	43,36%	40,92%
Alto (325-500)	65,96%	18,06%	5,21%
Total	100%	100%	100%

*Tabla 20 Porcentaje de estudiantes por Puntaje Global Pruebas Saber 11*

De allí se observa que:

1. El 65.96% de los estudiantes de los colegios privados bilingües tiene un puntaje alto y solo el 5.21% de los estudiantes de colegios públicos se encuentran en el mismo nivel de desempeño.
2. Mientras que el 92.47% de los estudiantes de colegios privados bilingües tiene puntaje entre medio alto y alto, mientras que el 53.87% de los estudiantes de colegios públicos solo están en bajo y medio.

A continuación, se realizó la siguiente prueba de hipótesis en RStudio, para su validación:

1. Prueba de Hipótesis 1

```

##### Prueba de hipótesis #####
## Más del 65% de los estudiantes de los colegios privados bilingües tiene un promedio alto >= 325:
## ?? = 0.05
## H0: p = 65
## H1: p < 65

zPrB <- (NROW(subset(nooficialB, nooficialB$PUNT_GLOBAL>= 325))/NROW(nooficialB) - 0.65) /
sqrt(0.65 * (1 - 0.65) / NROW(nooficialB))
zPrB # Para obtener el valor del estadístico
#[1] 3.186901

pnorm(q=zPrB, lower.tail=TRUE) # Para obtener el valor-P
#[1] 0.999281

```

*Ilustración 24 Prueba de hipótesis 1 en RStudio privado*

```

133
134 ## Menos del 5% de los estudiantes de colegios públicos tiene un promedio alto >= 325:
135 ## ?? = 0.05
136 ## H0: p = 95
137 ## H1: p < 95
138
139 zPu <- (NROW(subset(oficial, oficial$PUNT_GLOBAL<= 325))/NROW(oficial) - 0.95) /
140 sqrt(0.95 * (1 - 0.95) / NROW(oficial))
141 zPu
142 #[1] 2.625129
143
144 pnorm(q=zPu, lower.tail=TRUE) # Para obtener el valor-P
145 #[1] 0.9956692
146
147
148

```

*Ilustración 25 Prueba de hipótesis 1 en RStudio publico*

Con el resultado obtenido, concluimos que no se encuentra información suficiente para rechazar ninguna de las dos hipótesis y aceptamos la hipótesis de que el 65% de los estudiantes de colegios privados bilingües tiene un puntaje en desempeño alto, o sea mayor o igual a 325 puntos y solo el 5% de los estudiantes de colegios públicos tienen el mismo desempeño.

## 2. Prueba hipótesis 2

```
154
155 ##### Prueba de hipótesis 2 #####
156
157 ## Más del 92% de los estudiantes de los colegios privados bilingües tiene un promedio
158 ## mayor a medio alto >= 251:
159 ## ?? = 0.05
160 ## H0: p = 92
161 ## H1: p < 92
162
163
164 zPrB <- (NROW(subset(noOficialB, noOficialB$PUNT_GLOBAL >= 251))/NROW(noOficialB) - 0.92)
165 / sqrt(0.92 * (1 - 0.92) / NROW(noOficialB))
166 zPrB # Para obtener el valor del estadístico
167 #[1] 1.810646
168
169 pnorm(q=zPrB, lower.tail=TRUE) # Para obtener el valor-P
170 #[1] 0.9649022
171
172
170:15 Prueba de hipótesis 2 R Script
```

*Ilustración 26 Prueba de hipótesis 2 en RStudio privado*

```
172 ## Mas del 53% de los estudiantes de colegios públicos tiene un promedio menor a medio
173 ## <= 250:
174 ## ?? = 0.05
175 ## H0: p = 53
176 ## H1: p < 53
177
178 zPu <- (NROW(subset(oficial, oficial$PUNT_GLOBAL <= 250))/NROW(oficial) - 0.53) /
179 sqrt(0.53 * (1 - 0.53) / NROW(oficial))
180 zPu
181 #[1] 50.74364
182
183 pnorm(q=zPu, lower.tail=TRUE) # Para obtener el valor-P
184 #[1] 1
185
186
170:15 Prueba de hipótesis 2 R Script
```

*Ilustración 27 Prueba de hipótesis 2 en RStudio publico*

Con el resultado obtenido, concluimos que no se encuentra información suficiente para rechazar ninguna de las dos hipótesis y aceptamos la hipótesis de que más del 92% de los estudiantes de colegios privados bilingües tiene un puntaje en desempeño medio alto o alto, es decir mayor o igual a 251 puntos y más del 53% de los estudiantes de colegios públicos están por debajo de desempeño medio es decir menos de los 250 puntos.

### 3.9 Analítica diagnóstica de datos sobre las variables de valor

Al observar las diferentes tablas de puntuaciones por áreas, discriminando la información, y teniendo en cuenta las variables de naturaleza del colegio y el carácter de bilingüismo<sup>13</sup>, encontramos que:

- Los colegios bilingües tienden a tener valores mayores a 0 como valor mínimo en las áreas.
- La única área en la que los colegios bilingües privados tienen un resultado en 0 es en el puntaje de inglés.
- Aunque se evidencia que el porcentaje más alto de los estudiantes de colegio bilingüe privado es en inglés, ya que un 50% de estos estudiantes se encuentran entre un puntaje de 82 a 100 puntos, la repercusión del inglés en el puntaje final es la menor, con 1/13 de proporcionalidad con respecto a las otras puntuaciones que representan 3/13 cada una. Indicando que, aunque inglés sí tiene un puntaje superior, la afectación de los otros puntajes son los que realmente afectan el puntaje global.

De la misma manera, al visualizar el porcentaje de estudiantes según su nivel<sup>14</sup> identificamos que:

- El mayor porcentaje de estudiante que pertenece al colegio privado bilingüe tiene una puntuación en desempeño Alto, mientras que los estudiantes de colegio público están en desempeño Medio.

---

<sup>13</sup> Tablas 5,6, 7, 8, 9 y 10

<sup>14</sup> Tabla 20

- En esta tabla evidenciamos que existe una afectación significativa entre pertenecer a un colegio privado bilingüe que a un colegio público.

## CAPÍTULO IV: CONCLUSIONES Y RECOMENDACIONES

### 4.1 Conclusiones

- Los datos utilizados en el presente trabajo fueron obtenidos a partir de bases de datos abiertas y gratuitas, lo que facilitó el acceso a la información. Existen muchos entes gubernamentales y no gubernamentales que comparten los datos recolectados para su uso por parte de la población.
- Se identifica que para este tipo de datos es de crucial importancia para manejar un buen proceso ETL, puesto que al faltar datos y al tener un formato origen que afecta el formato utilizado por los procesadores de datos, puede generar datos erróneos en el análisis.
- Existe una diferencia significativa entre la población de estudiantes de colegios privados bilingües y el resto, ya que tienden a tener mejores resultados en el puntaje global comparado con los estudiantes de otras categorías obtenidas dentro del estudio realizado.
- Se encontró que, en todas las categorías de las puntuaciones establecidas en las tablas anteriores, en todos los casos las puntuaciones de los colegios públicos y privados bilingües inician en un valor superior a cero, excepto en las puntuaciones de inglés.
- Con el análisis descriptivo logramos aproximarnos a un análisis diagnóstico e identificar que existe una influencia en el puntaje global con respecto a que el estudiante sea de un colegio privado bilingüe y un colegio público.
- Con el resultado que obtuvimos no se explica nada aún, a pesar de que se da un indicio, todavía hay mucho por hacer desde el rol del Científico de datos.

- Al desarrollar la tesis en el modelo de espiral, logramos identificar los factores que afectan los resultados de las pruebas Icfes Saber 11 y patrones de comportamiento entre las áreas, generando más preguntas y una materia prima para posteriores trabajos, tanto para profundizar en el tema con las variables de valor utilizadas como para complementar con otras variables e identificar otros elementos de incidencia.
- Los datos del presente estudio poseen un nivel de madurez en el ámbito de la analítica descriptiva y la analítica diagnóstica, por tanto, se considera prematuro aplicar algoritmos de Machine Learning o aprendizaje de máquina, propios de la analítica predictiva y la analítica prescriptiva.
- Identificamos que el comportamiento del puntaje de un área a otra no es significativo y tiende a comportarse de manera similar una con la otra, y el puntaje global sigue el patrón de las 5 áreas evaluadas.
- Finalmente encontramos que existe un vínculo entre el bilingüismo y la naturaleza del colegio, puesto que la población que principalmente se diferencia del resto son los estudiantes de colegio privado bilingüe.

## 4.2 Recomendaciones y Trabajos futuros

- Teniendo en cuenta el análisis descriptivo de los resultados, como trabajo futuro se podría implementar un modelado de análisis diagnóstico y/o predictivo utilizando Machine Learning.
- Desde el ámbito de la ingeniería de datos, se podría diseñar un software para automatizar el proceso de ETL de los datos.
- Se pueden enfocar un poco más las poblaciones para visualizar si existe o no influencia, no solo por naturaleza de colegio y el bilingüismo sino también por zonas geográficas.
- Se puede enfocar igualmente en el nivel socioeconómico como una variable de valor, donde podría identificar si esta variable afecta en el puntaje del estudiante.
- Finalmente es de resaltar que se pueden encontrar alrededor de 82 variables en las bases de datos originales, de las cuales existe un enorme camino que se puede desarrollar y que puede arrojar diferente información con variables tanto socioeconómicas, geográficas, familiares, escolares entre otras. En la búsqueda puede que se encuentren otros factores de influencia en los resultados de las pruebas Saber 11.

## CAPÍTULO V: REFERENCIA BIBLIOGRÁFICA

[1] ICFES. (2019, 6 13). 50 años de historia. 50 AÑOS DEL ICFES.

<https://www.icfes.gov.co/50-icfes>

[2] Orjuela, J. (2014, 3 1). Determinantes individuales de desempeño en las pruebas de Estado para educación media en Colombia. ICFES.

<https://www.icfes.gov.co/documents/20143/233983/Determinantes+individuales+desempeno+pruebas+estado+para+educacion+media+en+colombia.pdf>

[3] Abadía Alvarado, L. K., Bernal Acevedo, G. L., & Muñoz, S. (2018). Brechas en el desempeño escolar en PISA ¿Qué explica la diferencia de Colombia con Finlandia y Chile? Education Policy Analysis Archives, 26(82), 37. <https://epaa.asu.edu/ojs/article/view/3423>

[4] Chaves Restrepo, M. (2019, 12 13). Los colegios de calendario B obtuvieron mejores resultados en las pruebas de Estado. La República.

<https://www.larepublica.co/especiales/mejores-colegios-2019/los-mejores-colegios-para-2020-a-la-luz-de-los-resultados-de-las-pruebas-saber-2943847>

[5] ICFES. (2021). Saber 11. Acceso a Bases de datos y diccionarios.

<https://www.icfes.gov.co/investigadores-y-estudiantes-posgrado/acceso-a-bases-de-datos>

[6] Media, O. (2017). Big Data Now: 2016 Edition. O'Reilly Media, Inc. 978-1-491-97748-4

[7] Miner, D. (2016). Hadoop: What You Need to Know. O'Reilly Media, Inc. 978-1-491-93730-3.

[8] Banker, K. (2012). MongoDB in Action. Manning Pubs Co Series. 978-1-935-18287-0

- [9] Bradshaw, S., Chodorow, K., & Brazil, E. (2019). MongoDB: The Definitive Guide: Powerful and Scalable Data Storage. O'Reilly Media, Inc. 978-1-491-95446-1
- [10] Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill España. 978-8-448-13610-9
- [11] Pérez López, C. (2011). Técnicas de análisis multivariante de datos. Aplicaciones con SPSS. PEARSON EDUCACIÓN. 978-84-205-4104-4
- [12] Miloslavskaya, N., & Tolstoy, A. (2016). Big data, fast data and data lake concepts. Procedia Computer Science, 88, 300-305. S1877050916316957
- [13] Hai, R., Geisler, S., & Quix, C. (2016, June). Constance: An intelligent data lake system. In Proceedings of the 2016 International Conference on Management of Data. 2097-2100.  
<https://dl.acm.org/doi/10.1145/2882903.2899389>
- [14] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. 978-14-939-3843-8
- [15] Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press. 9780-262-3043-20
- [16] Tang, E., & Fan, Y. (2016, November 16-18). Performance comparison between five NoSQL databases. IEEE, In 2016 7th International Conference on Cloud Computing and Big Data (CCBD), 105-109. 10.1109/CCBD.2016.030
- [17] Louridas, P., & Ebert, C. (2016). Machine Learning. IEEE Software, 33(5), 110–115.  
<https://doi.org/10.1109/MS.2016.114>
- [18] Raschka, S., & Mirajalili, V. (2015). Python machine learning (1st ed.). Packt Publishing. 9781783555130

- [19] García, J., Berlanga, A., Patricio, M., & Padilla, W. (2018). Ciencia de datos. Técnicas Analíticas y Aprendizaje Estadístico. Publicaciones Altaria, SL. Bogotá, Colombia. 978-84-947319-6-9
- [20] Marqués Asensio, F. (2017). R en profundidad: programación, gráficos y estadística (1st ed.). Alfaomega Grupo Editor, S.A. 978-607-622-973-6
- [21] Caballero, R., Martín, E., & Riesco, A. (2019). Big Data con Python (1st ed.). ALFAOMEGA. 978-958-778-577-0
- [22] Selwyn, N. (2014, May 28). Data entry: Towards the critical study of digital data and education. *Learning, Media and Technology*, 40(1), 64-82. 17439884.2014.921628
- [23] L'heureux, A., Grolinger, K., Elyamany, H.F., & Capretz, M.A. (2017). Machine learning with big data: Challenges and approaches. *Computer Science*, 5. 7776-7797
- [24] Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., & Taha, K. (2015, 9 1). Efficient machine learning for big data: A review. *Big Data Research*, 2(3), 87-93. 1503.05296
- [25] Kowarik, A., & Templ, M. (2016, 10). Imputation with the R Package VIM. *Journal of Statistical Software*, 74(7), 1-16. 10.18637.
- [26] Tsai, C.W., Lai, C.F., Chao, H.C., & Vasilakos, A.V. (2015). Big data analytics: a survey. *Computer Science - Journal of Big data*, 2(1), 1-32. 10.1186/s40537-015-0030-3.
- [27] Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A., & Buyya, R. (2015, 08 22). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, 79, 3-15. 10.1016/j.jpdc.2014.08.003.

[28] Diouf, P.S., Boly, A., & Ndiaye, S. (2018, June 11). Variety of data in the ETL processes in the cloud: State of the art. IEEE International Conference on Innovative Research and Development (ICIRD), 1-5. 10.1109/ICIRD.2018.8376308.

[29] Campello, R.J., Moulavi, D., Zimek, A., & Sander, J. (2015, July). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 10(1), 1-51. 10.1145/2733381.

[30] Rao, R.V., & Selvamani, K. (2015, May 22). Data security challenges and its solutions in cloud computing. Procedia Computer Science, 48, 204-209. 10.1016/j.procs.2015.04.171.

[31] Instituto Colombiano para la Evaluación de la Educación - Icfes. (n.d.). Documentación del examen Saber 11.

[https://www.icfes.gov.co/documents/20143/1885630/1.+Documentacion\\_Saber11.pdf/e72d7e45-7b05-fbee-aed7-c0dfafa25e2f?t=1590543922537](https://www.icfes.gov.co/documents/20143/1885630/1.+Documentacion_Saber11.pdf/e72d7e45-7b05-fbee-aed7-c0dfafa25e2f?t=1590543922537)

[32] Sistema Nacional de Evaluación Estandarizada de la Educación. (n.d.). Alineación del examen SABER 11°.

<https://www.icfes.gov.co/documents/20143/193784/Alineacion%20examen%20Saber%2011.pdf>

[33] Ministerio de Educación Nacional. (2006, mayo). Estándares Básicos de Competencias en Lenguaje, Matemáticas, Ciencias y Ciudadanas. [https://www.mineducacion.gov.co/1621/articles-340021\\_recurso\\_1.pdf](https://www.mineducacion.gov.co/1621/articles-340021_recurso_1.pdf)

[34] Instituto Colombiano para la Evaluación de la Educación - Icfes. (n.d.). DICCIONARIO DE VARIABLES SABER 11° PERIODO 20142 - 20182.

<https://www.icfes.gov.co/documents/20143/1885630/5.+Diccionario+Saber11+2014-2+a+2018-2.pdf>

[35] Instituto Colombiano para la Evaluación de la Educación - Icfes. (n.d.). DICCIONARIO DE VARIABLES SABER 11° PERIODO 20191 a 20192.

<https://www.icfes.gov.co/documents/20143/1885630/6.+Diccionario+Saber11+2019-1+a+2019-2.pdf>

[36] Aldas Manzano, J., & Uriel Jimenez, E. (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo, SA. 978-8428-3296-99.

[37] Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt Publishing. 978-0-544-00269-2

[38] *Saber 11°*. (n.d.). Niveles de desempeño Prueba de Inglés.

<https://www.icfes.gov.co/documents/20143/1500084/Niveles+de+desempeno+prueba+de+ingles.pdf/795d37dc-e3ee-d037-b926-0fbfc49e8efc>

[39] Martín, E., & Caballero, R. (2015). *Las bases de Big Data*. LA CATARATA. 978-8490-9777-50

[40] Pérez Márquez, M. (2015). *BIG DATA - Técnicas, herramientas y aplicaciones*. RC Libros. 978-84-943055-5-9

[41] Cortéz Morales, R. (2006). Introducción al Análisis de Sistemas y la Ingeniería de Software. Universidad Estatal a Distancia. 9977-64-961-8

[42] AWS Amazon. (n.d.). ¿Qué es NoSQL? Bases de datos no relacionales.

<https://aws.amazon.com/es/nosql/>

[43] Medina, F., & Galván, M. (2007). Imputación de datos: teoría y práctica. Publicación de las Naciones Unidas. 978-92-1-323101-2

[44] Castro Romero, Alexander, & González Sanabria, Juan Sebastián, & Callejas Cuervo, Mauro (2012). Utilidad y funcionamiento de las bases de datos NoSQL. Facultad de Ingeniería, 21(33),21-32. [fecha de Consulta 27 de junio de 2021]. ISSN: 0121-1129. Disponible en: <https://www.redalyc.org/articulo.oa?id=413940772003>

[45] Álvarez Jareño, J. A., & Coll Serrano, V. (2018). “Científico de datos”, la profesión del presente. *MÉI*, 9(16), 113-119. 2173-1241