Análisis del desempeño en inglés de las ingenierías en el eje cafetero a partir de resultados históricos en las pruebas Saber Pro en el marco de la virtualidad

Daniel Vasquez Alvarez

Código 1088344121

Directora de proyecto

Ingeniera Luz Estela Valencia Ayala

Universidad Tecnológica de Pereira

Facultad de Ingenierías

Ingeniería de Sistemas y Computación

Pereira

2022

Agradecimientos

A mi mejor amigo que me soporto hasta en los últimos instantes y me ayudo en incontables noches

Para mi padre que no pudo ver su sueño hecho realidad

Contenido

Listado de tablas

Listado de graficas

Introducción

El bilingüismo en la época globalizada en la que vivimos genera oportunidades tanto en el ámbito laboral como en la búsqueda del conocimiento y en la comunicación; en Colombia el interés por el dominio de una segunda lengua se ha visto reflejado en múltiples políticas, que buscan mejorar el conocimiento y las habilidades entre estudiantes de educación básica-media y educación superior siendo el inglés el idioma preferido. Sin embargo, este propósito requiere puntos de referencia internacionales que permita observar fácilmente el nivel de bilingüismo alcanzado en el ámbito Nacional, para ello se realizan evaluaciones periódicas mediante pruebas estatales obligatorias, siendo estas las pruebas Saber 11 y Saber Pro.

En el año 2020 la enseñanza de la segunda lengua se enfrentó al desafío de la virtualidad producto de la pandemia del Covid-19. La virtualidad exigió nuevas didácticas de enseñanza y los factores sociales como el acceso a los sistemas de información (computadores, teléfonos celulares, redes sociales, entre otros) y la comunicación (conexión a internet, cobertura celular) se convirtieron en ejes centrales para la educación. Esta situación facilita el análisis de las variables sociodemográficas y las relaciones entre ellas, que tuvieron repercusión sobre los resultados de inglés en las pruebas Saber Pro.

En este proyecto el análisis de estas variables: localización, el acceso a los dispositivos tics y la conexión a internet determinaron diferencias en los resultados obtenidos frente a años pasados. Para el análisis de los datos se utilizó el análisis descriptivo comparativo sobre los resultados de las pruebas Saber Pro de los años 2018, 2019, 2020 y 2021 contemplado dentro de la metodología de minería de datos para la información de los estudiantes de las carreras de ingeniería en los departamentos de Risaralda, Caldas y Quindío

Planteamiento del problema

¿Cuál es la incidencia de la pandemia y la virtualidad en el desempeño del bilingüismo dentro de los programas de ingenierías en las universidades del eje cafetero en la categoría de bilingüismo con base en las pruebas Saber Pro 2018 – 2021?

Desde el Estado se encuentra el objetivo de incrementar el nivel de bilingüismo entre los graduados de carreras profesionales. “En el año 2014 se planteó el Programa Nacional de inglés 2015-2025 Colombia Very well (MEN, 2014) con la iniciativa de continuar con lo planteado en el 2013, en un establecimiento de metas específicas para el 2025 en las pruebas Saber Pro. Se establece el objetivo de tener al 30% de los estudiantes en el nivel B1 y un 25% en el nivel B2 para los estudiantes fuera de la licenciatura de inglés” (Ministerio de Educación Nacional, 2015).

Sin embargo, estas metas no se han alcanzado, esto es transversal en todas las ciudades, lo cual indica un problema de índole nacional no específico. Las condiciones, por lo tanto, de la educación que reciben los estudiantes de pregrado en todo el país no cumplen con los metas que se han establecido como mínimos, esto es tener un puntaje igual o superior a B1 en la escala del Marco Común Europeo de Referencia para las Lenguas (MCERL).

En una perspectiva globalizada, es de interés del gobierno, de las universidades y de los mismos estudiantes tener la capacidad de dominar una segunda lengua que le conceda una ventaja en el mercado laboral, permitiendo ampliar la frontera laboral y en el ámbito macroeconómico facilitar los procesos de inversión extranjera en el país (Sánchez, 2013). En general, es una habilidad extra que tiene ventaja para los graduados y, bajo esta lógica, existen suficientes incentivos para su aprendizaje.

Para el año 2020 el desafío ya presente en la educación de una segunda lengua a nivel nacional se le suma las cuarentenas y la virtualidad. Tales condiciones generan presiones adicionales aún más fuertes y la obligación de identificar oportunidades de mejora en la enseñanza del inglés se convierte en una necesidad, para poder cumplir con las metas establecidas y garantizarle al mercado laboral profesionales con formación bilingüe.

Adicionalmente, es pertinente determinar los factores que incrementaron o redujeron la incidencia de la pandemia en los resultados de inglés, ya sean sociales, demográficos y geográficos en el ámbito nacional. Esto con fin de conocer los aciertos y desaciertos que se presentaron al tomar las decisiones educativas en la época de la pandemia. Con base en esta información se plantea la siguiente pregunta:

Justificación

Se busca poder conocer el dominio del segundo idioma dentro de los estudiantes de las instituciones del eje cafetero, puesto que el manejo de esta abre grandes oportunidades a nivel laboral en el mercado globalizado a próximos profesionales, sino que también es un atractivo de inversión extranjera al tener capital humano capaz de afrontar el reto de comunicación global.

Las instituciones educativas y el Estado convertido el bilingüismo en un objetivo siendo este último promotor de diversos proyectos para el mejoramiento del aprendizaje es importante saber si estas metas planteadas se cumplen o no, esto sumando la situación de irregularidad que fue la pandemia donde el aprendizaje exigió nuevas didácticas de enseñanza y los factores sociales como el acceso a sistemas de la información y la comunicación se convirtieron en ejes centrales para la educación

conocer como influyo esta permitirá evidenciar relaciones de fortaleza o de debilidad que se presentan en el aprendizaje de una segunda lengua.

En el caso del Eje Cafetero, es importante destacar la presencia de un gran número de universidades, fundaciones e instituciones universitarias que ofrecen educación a nivel superior y a su vez un mercado laboral en crecimiento con miras al sector servicio internacional. Esto genera una presión adicional en el mercado hacia la formación bilingüe, donde se hace necesario analizar si se está cumpliendo esta condición en los graduados de la región o no.

Objetivo general

Analizar, mediante un modelo descriptivo-comparativo, la incidencia de la pandemia y la virtualidad en el desempeño del bilingüismo de los programas de ingenierías de las universidades del eje cafetero en las pruebas Saber Pro 2018 – 2021.

Objetivos específicos

Identificar las variables centrales de análisis de los resultados de las pruebas Saber Pro 2018-2021, a partir de la recopilación, limpieza y transformación de los datos.

Analizar de forma descriptiva las variables identificadas, a través del uso de herramientas de análisis y modelos de prueba que permiten la comprobación de hipótesis.

Concluir sobre la incidencia que presentó la pandemia en las pruebas Saber Pro a partir de los resultados obtenidos.

Marco Teórico referencial

Educación de una segunda lengua

Desde diferentes gobiernos se han impulsado programas para impulsar el nivel de bilingüismo de la sociedad, generalmente, un bilingüismo enfocado en el inglés. Este interés particular tiene su origen en la importancia del inglés para la vida profesional y en menor medida cultural de la ciudadanía (Merchán, J. et al., 2021). En este sentido, es importante entender que se considera bilingüismo, sus indicadores, las estrategias de enseñanza y la adaptabilidad a un entorno virtual.

Bermúdez & Fandiño (2016) compilan las diferentes definiciones de bilingüismo a través de la teoría, las cuales se remontan desde Bloomfield (1933) donde se define como la habilidad de hablar dos lenguas con un dominio similar al nativo, hasta Lam (2001) como un fenómeno de competencias y comunicación en dos lenguas. Estas definiciones son similares, pero con implicaciones diferentes. Se entiende actualmente el bilingüismo como; la necesidad de unas competencias (básicas, medias o avanzadas) que la persona debe cumplir.

A su vez, estás competencias se pueden clasificar desde otras áreas como la edad a la que una persona las adquiere o el nivel de integración que hay entre ambas lenguas. Es a partir de esa interacción que el bilingüismo puede ir más allá y centrarse en el intercambio no solo de idiomas sino también de culturas o incluso la formación, a través del Multilingüismo, Plurilingüismo, Biculturalismo, Interculturalidad y Alfabetismo múltiple. (Bermúdez, Fandiño & Ramírez, 2014).

Si se vuelven a revisar las intenciones gubernamentales de bilingüismo es posible encasillarlo en un concepto de interculturalidad, donde se reconoce y acepta la diversidad cultural en un contexto comunicativo, social y de negocios entre dos grupos culturales, en este caso, con el mundo angloparlante (Reyes, 2011). Esta vinculación cultural se basa en la comunicación a través del

acceso a información en una segunda lengua, en este caso, el inglés. Sin embargo, para que se de este bilingüismo es importante entender las metodologías de aprendizaje de una lengua.

El aprendizaje de una segunda lengua se puede abordar desde diferentes métodos, estos incluyen: gramática-traducción, directo, audio-lingüístico, comunicativo y estructural (Benati, 2018); sin embargo, existen métodos aún más complejos enfocados en actividades particulares como Respuesta Física Total (TPR por sus siglas en inglés) en donde los estudiantes deben seguir instrucciones físicas vinculando el movimiento al aprendizaje (Bancroft, 1999); Enseñanza Comunicativa del Lenguaje (CLT por sus siglas en inglés) donde se entrelazan las estructuras del lenguaje con las perspectivas comunicativas del mismo (Richards, 2005); entre otros.

Finalmente, es posible observar un método denominado Aprendizaje de Lenguaje Asistido por Computadora (CALL por sus siglas en inglés) que toma especial interés por el uso de pantallas digitales y métodos de inmersión virtual para fortalecer el aprendizaje de los idiomas (Tafazoli, et al, 2019). Es importante destacar que esta metodología es difícil de revisar dado la constante evolución que se presenta, evolucionando tanto el apartado tecnológico como la didáctica inherente al mismo.

La adaptación a modelos de aprendizaje CALL parte del principio de la normalización del aprendizaje como parte de la cotidianidad digital del estudiante (Beatty, 2013). Para esto es necesario que se desarrollen las estrategias necesarias para que la aproximación al aprendizaje sea tanto educativo como didáctico. Envolviendo herramientas de otros métodos en un macro-método de aprendizaje que requiere de tiempo para ser implementado de forma exitosa (Utami, 2020).

Dada la novedad e imprevisibilidad de una situación como la pandemia del covid-19, la bibliografía recién empieza su exploración desde la aplicación de un método CALL obligatorio, que no logra

estar cimentado de inmediato, en especial en espacios de baja conectividad y que requiere de tiempo para su adaptabilidad. Por lo tanto, para evidenciar una recuperación del proceso de enseñanza, mediante el cambio de estrategias docentes, donde no hay un reemplazo de la educación cara a cara, pero sí una solución temporal y, más importante, adaptable en situaciones como una cuarentena (Segura et al, 2020).

En este sentido, el acceso de las herramientas de la Tecnología de la Información y la Comunicación (TIC) son centrales para el digital. No obstante, como herramientas tienen aspectos tanto positivos como negativos, que incluyen facilidad para el plagio, poco o ningún interés en los sistemas o plataformas y, por supuesto, el fácil o difícil el acceso a internet o dispositivos de cómputo para el aprendizaje (Fansury et al, 2020). Como se mencionaba con las estrategias de aprendizaje CALL la adaptación de estas estrategias es algo que requiere tiempo y entrenamiento por parte de los enseñantes.

La cuarentena y la pandemia no permitieron el tiempo de estabilización necesario y, por lo tanto, se espera que los resultados no hayan sido los mejores dentro del rango esperado de aprendizaje digital; sin embargo, es esperable que los aprendizajes en materia de enseñanza fortalezcan las estrategias y se delimiten los aspectos que no están funcionando, teniendo en cuenta lo impactante de una situación no prevista que se permea de aprendizaje basado en página web, aprendizaje basado en sistema de cómputo y la interacción remota de los docentes (Gallegos, 2021).

En la adaptación a modelos de aprendizaje digitales para una segunda lengua es importante tener en consideración la preparación: tecnológica (empezando por el acceso a internet), de contenido (accesibilidad a los programas de forma virtual), de aprendizaje (aproximaciones pedagógicas y vinculativas de enseñanza) y de monitoreo (reunir resultados y adaptar procesos a partir de la información), esta preparación requiere de un análisis de la meta a alcanzar en conjunto con el

grupo poblacional y las herramientas disponibles, situación que un evento particular como el covid-19 no facilitó (Zorcic, 2020).

La minería de datos CRISP-DM

Es, precisamente, en el apartado de monitoreo que es importante aplicar estrategias de análisis con base en la información. Establecer herramientas comparativas que permitan entender los aspectos que se han abordado de forma exitosa y aquellos que tienen oportunidades de mejora. Sin embargo, es necesario que estos datos existan y que puedan ser recolectados para que se pueda darle tratamiento para su análisis.

El análisis comienza a través de mecanismos de minería de datos, siendo este un proceso para extraer conocimiento útil y de actualidad de una gran colección de datos, en busca de encontrar relaciones e inferencias que no se pueden obtener con métodos estadísticos regulares (Moine, 2012). Esto toma relevancia por la gran cantidad de variables que se interconectan en la revisión de avances y retrocesos del aprendizaje y sus respectivos programas.

Para alcanzar estos objetivos se presentan diferentes metodologías que se engloban en un Proceso de Descubrimiento de Conocimiento (KDD por sus siglas en inglés), el cual se define como un "proceso de extracción no trivial de información implícita (..) potencialmente útil de datos" (Frawley, Piatetsky & Matheus, 1992).

Este proceso, para su facilidad e implementación se descompone en fases que permiten la manipulación de la información desde los datos iniciales, pasando por su almacenaje, selección, la identificación de patrones (siendo este el apartado donde se implementa la minería de datos), la generación de conocimiento y, su presentación para la toma de decisiones (Larose & Larose, 2014).

En otras palabras, descomponer y ajustar la información en grandes porciones de datos para poder realizar análisis y construir resultados a partir de tal información.

El método de análisis principal es el Proceso Estandarizado Cross Industrial para Minería Datos CRISP-DM (por sus siglas en inglés). Este es un método que se comenzó a utilizar a finales del siglo XX como un proceso jerárquico basado en tareas como línea base para la manipulación de datos en un proceso de minería (Khasanah & Harwati, 2019). Lo que realiza este método es poner a disposición un método estandarizado para que los datos minados se ajusten a las estrategias de solución de problemas para la investigación y los negocios.

Schröer, Kruse & Marx (2021) descomponen las 6 fases del modelo para explicar las aproximaciones que debe tener cada fase como una subsección indispensable para la investigación académica. Esto es importante ya que el modelo CRISP-DM no es únicamente un instructivo de modelación de datos, se concentra en la información, los datos que se reciben, los objetivos explícitos que se presentan y como los datos van a responder ese objetivo.

Una sección por cada fase de CRISP-DM	Métodos y enfoques como subsecciones
1. Entendimiento del negocio	1.1 Descripción textual en la sección propia 1.2 Definicion explicita del objetivo de la mineria de datos
2. Entendimiento de los datos	2.1 Mención de la fuente de datos y del proceso de recolección 2.2 Descripción estructural (modelo de datos, datos de ejemplo) 2.3 Estadistica descriptiva obligatoria
3. Preparacion de los datos	3.1 Describir los datos de entrada y salida 3.2 Métodos y enfoques (transformación, selección, limpieza)
4. Modelado	4.1 Mención del enfoque de modelización 4.2 Al menos la tecnologia utilizada deberia ser mencionada aqui 4.3 Construcción de conjuntos de prueba y de entrenamiento
5. Evaluacion	5.1 Definir las métricas 5.2 Visualización de los modelos y las métricas
6. Despliegue	6.1 Si el despliegue en el ámbito de aplicación, las implementaciones deben ser descritas

Tabla 1 Plantilla para la presentación y estructuración de los trabajos según CRISP-DM. Adaptado de Schröer, Kruse & Marx (2021)

1. Entendimiento del negocio: es la primera fase del proceso donde se realiza un entendimiento de los objetivos del proyecto y sus requerimientos desde una perspectiva del negocio, esto para posteriormente transformar este entendimiento en una definición de problema y un trazo del plan que se va a seguir en pos de alcanzar estos objetivos
2. Comprensión de los datos: en esta segunda fase empieza con el vistazo inicial de los datos y las actividades de entendimiento y familiarización de estos, donde se identifican los problemas que puedan tener, las percepciones de estos y donde las primeras hipótesis y se detectan los subconjuntos sobre los cuales se va a trabajar
3. Preparación de los datos: la fase de preparación de los datos es donde se abarca la construcción del conjunto de datos final a partir de los datos recopilados en bruto (sin ningún tipo de preprocesamiento)
4. Modelado: en esta fase es donde se seleccionarán las técnicas de modelización en donde se busca calibrar la base de datos obtenida y en donde los parámetros finales son escogidos de acuerdo con los requerimientos anteriormente planteados.
5. Evaluación: es la fase del proyecto donde evalúan los modelos obtenidos desde el punto de vista del análisis de datos, se evalúan las variables planteadas con el fin de asegurarse de que alcanza los objetivos planteados y llegado el caso determinar si hay alguna cuestión importante que no se haya tenido suficientemente en cuenta, el final de esta fase es donde se decide sobre el uso de los resultados del proceso de minería de datos
6. Despliegue: es la fase final del proyecto donde el conocimiento obtenido tendrá que organizarse y presentarse debidamente de una forma en que el cliente pueda utilizarlo

   (Chapman et al, 2000)

Desarrollo del proyecto

Identificar las variables centrales de análisis de los resultados de las pruebas Saber Pro 2018-2021, a partir de la recopilación, limpieza y transformación de los datos.

Las pruebas Saber Pro antes llamadas pruebas ECAES surgen a finales de la década de 1990, empiezan a plantearse como modelo de evaluación de la educación superior pero no es hasta el 2003 donde estas se reglamentan con aplicación inmediata. En el 2009 por debido a la ley 1324 y al decreto 3963 del 2009[1] estas cambian su nombre a Saber Pro, obligatorias para los estudiantes universitarios y se refuerza como instrumento de medida de la calidad de la educación superior nacional.

Su ventaja radica en la estandarización de las áreas fundamentales y específicas del saber. Esta cualidad facilita la generalización de resultados y permite al ICFES, las universidades y los investigadores realizar análisis a partir de la información. La obligatoriedad de los exámenes para los estudiantes próximos a graduarse aporta a las instituciones de educación superior realizar procesos de autoevaluación y autogestión de los contenidos programáticos, así como el pensum de sus programas académicos en busca de mayor calidad académica.

La prueba se categoriza en módulos de competencias con una distinción entre conocimientos generales y conocimientos específicos. Las competencias genéricas incluyen los siguientes cinco módulos: razonamiento cuantitativo, lectura crítica, competencia ciudadana, comunicación escrita e inglés. El ICFES define cada módulo elementos básicos para los profesionales a nivel nacional.

En el caso del inglés como segunda lengua de interés nacional tiene un peso particular debido al interés gubernamental en incrementar el dominio en esta lengua, especialmente entre los profesionales nacionales, en pro de incrementar la competitividad en el mercado internacional, el acceso a fuentes de conocimiento, entre otros. Este objetivo incluye políticas para el mejoramiento incrementadas desde el año 2014 en el

[1] Documentación del examen Saber Pro ICFES

Programa Nacional de inglés 2015 – 2025 Colombia Very Well (MEN, 2014) basadas en el cumplimiento de metas y teniendo como variable objetivo el puntaje obtenido en las pruebas Saber Pro.

El ICFES publica los resultados de la prueba Saber Pro de forma anualizada tanto a nivel individual para los estudiantes como un compilado general. Esta información es de carácter público tanto el año actual como los años anteriores disponible en el portal virtual del ICFES[2]. A su vez, gracias a la información sociodemográfica que se recolecta de cada participante es posible categorizar los resultados, según: sexo, nacionalidad, estrato, lugar de estudio, lugar de residencia, presencia de discapacidades, valor de la matrícula, universidad y facultad, entre otras condiciones que pueden ser un factor de cambio en los resultados.

A través de la información suministrada por el ICFES se encuentran los resultados a las pruebas Saber Pro de los años 2018 al 2021. De estos cuatro años se tiene tanto los valores obtenidos por los estudiantes en cada módulo, así como las respuestas a las preguntas sociodemográficas que permiten caracterizar los estudiantes y realizar procesos de inferencia. Debido a la cantidad de información se utiliza un proceso de depuración de información, el cual limita la cantidad de datos a 7368 con 26 variables de análisis.

Teniendo en cuenta la cantidad de datos con la que se cuenta es necesario tener una metodología para poder trabajar de forma efectiva con ellos, se tomó un enfoque simplificado de los métodos de minería de datos de CRISP-DM orientado a la creación de un producto de datos siendo este el resultado de la explotación de los datos donde se contemplan los siguientes pasos de análisis de este:

Comprensión del negocio: la primera fase está en el entendimiento de las pruebas Saber Pro, inicialmente desde un enfoque histórico y posteriormente desde el punto de vista del componente central de la investigación, la prueba de inglés, con el apoyo de la documentación proporcionada por el ICFES, se

[2] Base de datos del icfes (Data icfes)

establece el contexto y los objetivos tempranos del análisis a realizar complementado con información sobre la importancia del inglés para los recién graduados.

Comprensión de los datos: la segunda fase está en la recopilación de los datos necesarios de la base de datos que posee el ICFES para las pruebas Saber Pro identificando la calidad de estos como también estableciendo las relaciones tempranas que permitan llevar a hipótesis de las variables, esto permite entender como están distribuidos los datos y que procesos aplicar sobre ellos .Se seleccionan los resultados desde el 2018 al 2021 y a partir de la explicación del contenido de los datos se seleccionan las variables potenciales para el análisis.

Preparación de los datos la tercera fase identifica los datos se realizó un proceso de extracción de los datos de la base del ICFES donde para el set de datos escogidos la cual posee 107 campos; para hacer el primer tratamiento sobre los datos, primero retiramos todos los campos redundantes, generamos variables nuevas que nos faciliten la identificación de cada año, cambios de formato en ciertas variables y eliminación de campos mayoritariamente vacíos o que no tengan algún tipo de valor para el análisis, por lo cual el primer conjunto de datos está conformado por 26 (Tabla 2) variables usando así el 24,3% de los datos descartando así un 75,3 % del conjunto original, Teniendo un primer conjunto de datos se unen las 4 bases de datos en una sola para integrarlas a una base de datos no relacional de manejo eficiente.

Nombre Campo	Nombre variable	Tipo
Tipo de documento	ESTU_TIPODOCUMENTO	Texto
Nacionalidad	ESTU_NACIONALIDAD	Texto
Genero	ESTU_GENERO	Texto
Departamento de residencia del estudiante	ESTU_DEPTO_RESIDE	Texto
Municipio de residencia del estudiante	ESTU_MCPIO_RESIDE	Texto
Área de residencia del estudiante (cabecera municipal o área rural)	ESTU_AREARESIDE	Texto
Valor de matrícula universidad	ESTU_VALORMATRICULAUNIVERSIDAD	Texto
Estrato de la vivienda	FAMI_ESTRATOVIVIENDA	Texto
Tiene internet (si o no)	FAMI_TIENEINTERNET	Texto

Tiene computador (si o no)	FAMI_TIENECOMPUTADOR	Texto
Nombre de la institución	INST_NOMBRE_INSTITUCION	Texto
Programa académico	ESTU_PRGM_ACADEMICO	Texto
Grupo de referencia	GRUPOREFERENCIA	Texto
Núcleo del pregrado	ESTU_NUCLEO_PREGRADO	Texto
Departamento de la institución	ESTU_INST_DEPARTAMENTO	Texto
Carácter académico de la institución	INST_CARACTER_ACADEMICO	Texto
Origen de la institución	INST_ORIGEN	Texto
Puntaje del módulo razonamiento cuantitativo	MOD_RAZONA_CUANTITAT_PUNT	Numérico
Puntaje del módulo lectura critica	MOD_LECTURA_CRITICA_PUNT	Numérico
Puntaje del módulo competencia ciudadana	MOD_COMPETEN_CIUDADA_PUNT	Numérico
Puntaje del módulo ingles	MOD_INGLES_PUNT	Numérico
Clasificación del desempeño de ingles del estudiante	MOD_INGLES_DESEM	Texto
Puntaje del módulo comunicación escrita	MOD_COMUNI_ESCRITA_PUNT	Numérico
Puntaje global	PUNT_GLOBAL	Numérico
Percentil global	PERCENTIL_GLOBAL	Numérico
Año	Anio	Numérico

Tabla 2 datos preprocesados

Modelado En esta cuarta fase después de realizar la preparación de los datos previa se realiza un último filtro sobre los datos ya transformados para abordar solo los departamentos de interés: Risaralda, Quindío y Caldas. Igualmente, el análisis se centra en el núcleo académico de las ingenierías donde de esta, se escogieron 12 campos para realizar el análisis lo cual corresponde a un 46,15% de los datos preprocesados anteriormente y a un 11,21% de los datos totales.

Nombre del campo	Nombre de la variable	Valoración
Tipo de documento	ESTU_TIPODOCUMENTO	1
Área de residencia del estudiante (cabecera municipal o área rural)	ESTU_AREARESIDE	4
Tiene internet (si o no)	FAMI_TIENEINTERNET	5

Tiene computador (si o no)	FAMI_TIENECOMPUTADOR	5
Programa académico	ESTU_PRGM_ACADEMICO	1
Núcleo del pregrado	ESTU_NUCLEO_PREGRADO	1
Departamento de la institución	ESTU_INST_DEPARTAMENTO	5
Origen de la institución	INST_ORIGEN	3
Puntaje del módulo ingles	MOD_INGLES_PUNT	4
Clasificación del desempeño de ingles del estudiante	MOD_INGLES_DESEM	5
Percentil global	PERCENTIL_GLOBAL	2
Puntaje global	PUNT_GLOBAL	3
Año	Anio	4

Tabla 3 Jerarquía de los datos procesados

Teniendo así la base de datos final limpia y transformada con los datos preparados se implementa un modelo donde se califica la importancia de los campos preseleccionados siendo 5 la calificación más alta y 1 la calificación más baja de esta forma se diseña el producto de datos el cual apunta a un análisis descriptivo en el cual se tiene entendimiento sobre como interactúan las variables, la influencia que tienen en el modelo y la representación que permita validar una prueba de hipótesis y concluir (Tabla 3).

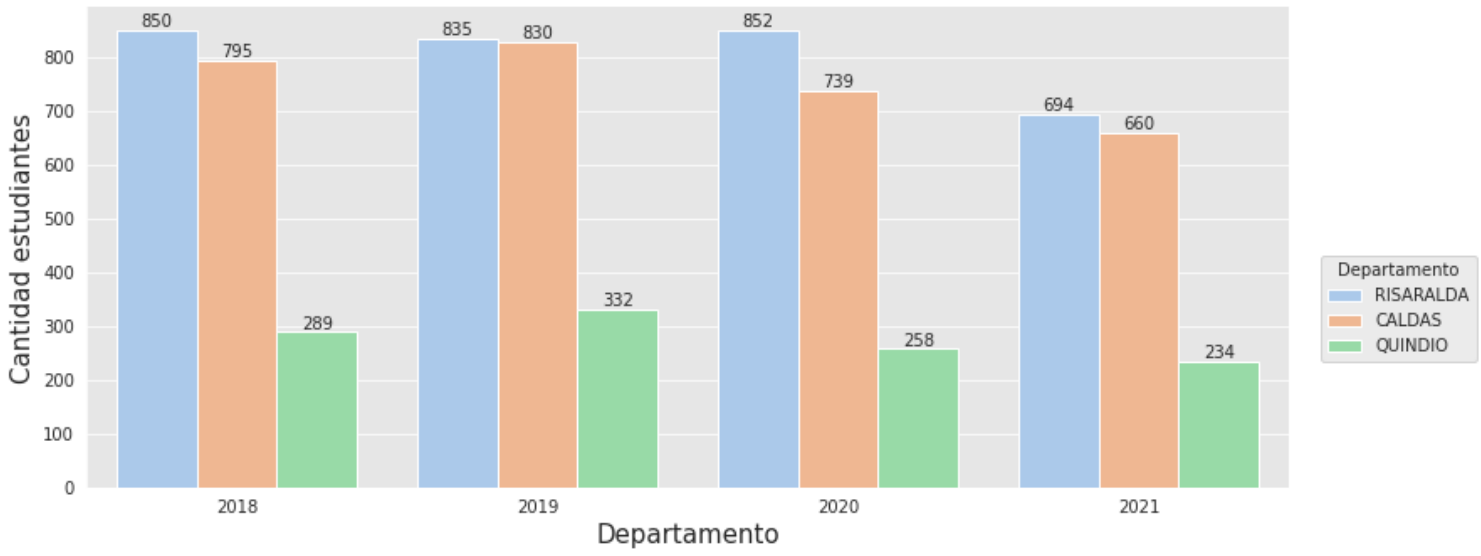
Evaluación esta es la quinta fase en donde se explora los datos y se busca establecer las necesidades establecidas en la investigación la cual se explora a continuación.

Analizar de forma descriptiva las variables identificadas, a través del uso de herramientas de análisis y modelos de prueba que permiten la comprobación de hipótesis

Para los intereses de la investigación es importante separar los grupos en subconjuntos que permiten realizar comparaciones entre ellos. Estos subconjuntos tienen dos filtros, el año de presentación de la prueba y el departamento.

Para el año 2018, 1.934 estudiantes presentaron la prueba, 1.997, 1.849 y 1.588 fueron los datos del estudiante respectivamente para los años 2019, 2020 y 2021, esto ubica el 2021 como el año donde mayor deserción o aplazamientos se presentaron. De estos totales a su vez es posible desagregar la información como se presenta en la Gráfica 1 estableciendo la cantidad de estudiantes que presentaron su prueba Saber Pro en cada departamento por año.

Lo anterior brinda dos datos importantes, en primer lugar, el peso total de cada departamento con respecto a sus estudiantes, de forma generalizada, Risaralda y Caldas tienen una mayor cantidad de individuos frente a Quindío sin importar el año que se analice.



Gráfica 1 Distribución de estudiantes por año en los departamentos

En segundo lugar, es posible ver la evolución del total de estudiantes que presentaron las pruebas Saber Pro en cada año por departamento y, por consiguiente, hacer inferencia con respecto a cuáles se vieron más

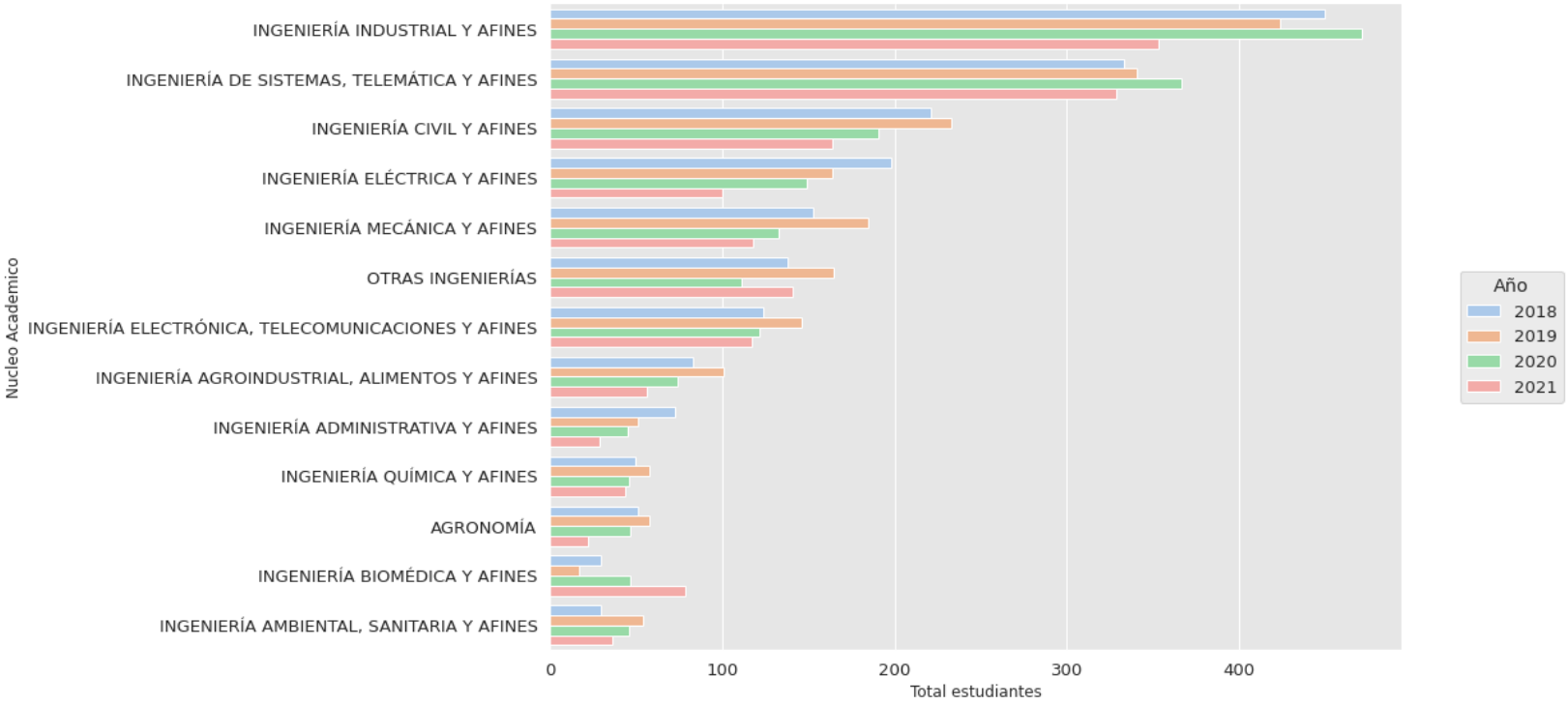
afectados por la pandemia del COVID en cuanto a cantidad como tal, sin entrar en el análisis de los puntajes en una primera instancia. Por eso es importante notar que del año 2018 al 2019 se presenta un crecimiento de estudiantes para Caldas y Quindío, sin embargo, para el 2020 Risaralda es el único departamento que incrementa el número de estudiantes que presentan la prueba Saber Pro, en contraposición a la disminución presente durante el año 2019 donde este fue el único de los tres departamentos con este comportamiento. Lo anterior permite inferir que se presentaron políticas o estrategias diferentes en el departamento de Risaralda en el 2020 que permitieron un crecimiento cuando los otros departamentos disminuyeron su participación. En síntesis, Risaralda creció para el 2020 un 2.0% interanual, mientras Caldas presenta una disminución del 11% y Quindío una del 22.2%.

Finalmente, el año 2021 termina con un comportamiento recesivo para los tres departamentos , dónde incluso Risaralda que había preservado una leve tendencia positiva cae muy por debajo de los niveles del año 2018; con un 18.5% , siendo, sin duda, el principal afectado; El departamento de Caldas presento una disminución del 10.7% y Quindío con una del 9.3% Este comportamiento es coherente con la disminución de matrículas estudiantiles en las universidades para el último semestre del 2020 y primero del 2021.

Esta información ilustra el comportamiento dentro de los tres departamentos en el cual se presenta un elevado nivel de deserción o aplazamiento de la carrera universitaria, y, por lo tanto, el primer impacto que se puede percibir a causa de la pandemia es una disminución de la cantidad de graduados, considerando las pruebas Saber Pro como un requisito para la culminación de la carrera universitaria.

Para ampliar aún más el contexto con respecto a la incidencia de la pandemia en la cantidad de estudiantes que presentan las pruebas Saber Pro siendo estos próximos a graduarse, es pertinente revisar este comportamiento de la deserción en los diferentes núcleos académicos disponibles en los años seleccionados dentro de la categoría de “Ingenierías” (Gráfica 2). Se destacan de forma particular las ingenierías de sistemas y telemática, las ingenierías industriales y las ingenierías civiles como los núcleos dónde más estudiantes presentan las pruebas de manera consecutiva en todos los años.

Para la ingeniería en sistemas y telemática junto con la ingeniería industrial los valores en el año de la pandemia 2020 fue donde más estudiantes presentaron la prueba en cada ingeniería y donde finalmente en el 2021 ingeniería de sistemas que tuvo una cantidad cercana a la del 2018 siendo 329 y 333 respectivamente, mientras que la ingeniería industrial tuvo un decrecimiento en el 2021 de casi 100 estudiantes, es decir un 25.0% donde esta ingeniería presenta la menor cantidad de estudiantes dentro de los cuatro años del estudio, la ingeniería civil es otra ingeniería junto a sistemas e industrial que tenía una alta cantidad de estudiantes para los años 2018 y 2019 pero la pandemia redujo esta cantidad de estudiantes en el 2020 un 18% y en el 2021 un 14.1%. Los comportamientos de crecimiento y decrecimiento entre años son similares para los otros once núcleos de ingeniería, gran parte tienen crecimiento en los años previos a la pandemia y sufren caídas en el número de estudiantes durante la pandemia siendo así que para el último año los valores son generalmente inferiores a los períodos anteriores.



Gráfica 2 Distribución de núcleos académicos por año

Dentro del análisis de la información, se encuentran similitudes y diferencias dentro de la composición del ambiente universitario de los tres departamentos analizados. El punto de comparación es la cantidad de estudiantes según el tipo de institución de educación superior que se presentaron las pruebas Saber Pro en los años estudiados, considerando que para Risaralda y Caldas el grueso de los estudiantes se encuentra en universidades de carácter oficial – nacional, la mayoría de los estudiantes del Quindío estudian en una institución oficial – departamental (Tabla 4).

INST_ORIGEN	CALDAS	QUINDIO	RISARALDA
NO OFICIAL - CORPORACIÓN	561	86	606
NO OFICIAL - FUNDACIÓN	424	291	133
OFICIAL DEPARTAMENTAL	-	736	-
OFICIAL NACIONAL	2039	-	2492

Tabla 4 Distribución de estudiantes por tipo de institución

Lo anterior se explica debido a la oferta particular de instituciones de educación superior en el territorio. Dado que la oferta no es muy amplia, los estudiantes se concentran y se distribuyen como muestra en la Tabla 5. En este sentido se destacan cuatro instituciones de forma principal, en Risaralda la Universidad Tecnológica de Pereira, en el Quindío la Universidad del Quindío y en Caldas tanto la Universidad Nacional de Manizales y la Universidad de Caldas. Sin embargo, no son las únicas instituciones proveedoras de servicios de educación superior en la región también contando con universidades más pequeñas que ofrecen los programas de ingeniería dando así posibilidades de elección para el estudiante.

INST_ORIGEN	CALDAS	QUINDIO	RISARALDA
NO OFICIAL - CORPORACIÓN	1	2	3
NO OFICIAL - FUNDACIÓN	2	1	1
OFICIAL DEPARTAMENTAL	-	1	-
OFICIAL NACIONAL	2	-	1

Tabla 5 distribución de instituciones por departamento

Entrando en los factores sociales de la población se debe realizar una distinción que permitirá comparar la incidencia de la pandemia para los diferentes estudiantes. Los principales factores para analizar es la presencia de internet en el hogar para el estudiante, la presencia de un computador y si el estudiante habita en el área rural o cabecera municipal.

Estas tres variables son centrales para el análisis de la pandemia como factor de afectación del desempeño en el nivel de inglés de los estudiantes por cuenta de los efectos de la pandemia. El principio de análisis es simple, si la universidad presencial no estuvo en funcionamiento, los estudiantes sin acceso a internet deberían verse más afectados que aquellos que si tienen acceso, lo que generó un cambio que obliga a los estudiantes a adquirir una conexión a internet para asistir a sus clases.

Igualmente, el análisis con respecto al área de residencia de los estudiantes es más difícil de abordar, ya que su incidencia no está necesariamente vinculada a la afectación de la pandemia. Sin embargo, es posible plantear como hipótesis que los estudiantes de las zonas rurales con acceso a internet se vieron beneficiados al disminuir su tiempo y esfuerzo de desplazamiento, mientras los estudiantes de zonas rurales sin acceso a internet podrían verse perjudicados pues sus limitaciones eran mayores.

Acceso a Internet:

Aquellos quienes tienen internet y computador

Es importante saber algunos factores externos que pueden llegar a afectar los puntajes de inglés durante el periodo de la pandemia, factores como la conectividad a la red y el acceso a los medios de información en este caso el computador, esta información demográfica se recopila a través de los cuestionarios al previos del examen. Entre esas preguntas se encuentran dos condiciones la tenencia de computador y la tenencia de internet. Estos factores son fundamentales en el desempeño del módulo de inglés de la prueba debido a las condiciones de virtualidad que la pandemia origino donde la necesidad de ambos eran requisitos obligatorios para afrontar la virtualidad.

	Anio	2018		2019		2020		2021	
	FAMI_TIENECOMPUTADOR	No	Si	No	Si	No	Si	No	Si
FAMI_TIENEINTERNET	ESTU_INST_DEPARTAMENTO								
No	CALDAS	25	62	29	55	7	31	8	23
	QUINDIO	5	19	15	20	1	7	5	6
	RISARALDA	11	44	28	55	4	40	12	20
Si	CALDAS	12	643	17	707	17	671	12	575
	QUINDIO	3	243	5	277	6	238	9	202
	RISARALDA	25	724	23	690	14	779	18	608

Tabla 6. Cantidad de estudiantes que presentaron las Pruebas Saber Pro dependiendo de sus condiciones

En este sentido, analizando la situación antes de pandemia (2018 – 2019) para los tres departamentos se encuentra que aquellos estudiantes con conexión a internet y computador teniendo en cuenta la cantidad de estudiantes para el departamento de Caldas siendo de 643 para el 2018 y 707 para el 2019, de igual forma para Risaralda se registra para el 2018 724 y en el 2019 690, mientras para el departamento del Quindío 243 para el 2018 y 277 para el 2019 (Tabla 6) siendo de los tres departamentos el de menor cantidad de estudiantes que presentaron la prueba para ambos años, donde Caldas y Risaralda tienen promedios de nota superiores a los 160 puntos y en el 2019 por encima de los 165 puntos (Gráfica 3). Los cambios interanuales para este grupo poblacional son mínimos con incrementos del orden del 0.54% para Risaralda y 2.85% para Caldas y una disminución del 1.61% para Quindío (Tabla 7).

Entrando en el primer año de pandemia es importante notar un cambio en el puntaje positivo para Caldas y para Quindío de 3.73% y 3.61% respectivamente (Tabla 7); es importante contrastar estos datos con un menor número de individuos presentando las pruebas para ambos departamentos (Tabla 6). Sin embargo, también deben considerarse factores como la adaptabilidad a la virtualidad o las estrategias de educación determinadas en las primeras etapas de pandemia. A diferencia del departamento de Risaralda que presentó una disminución del 1.1%, con una condición diferencial a los otros dos departamentos y es la sostenibilidad y aumento en el número de estudiantes presentando las pruebas en este año.

Para analizar estos factores es necesario entender tanto las políticas que tomaron cada uno de los departamentos en materia educativa, en algunos casos entrando en las decisiones de cada universidad que podrían haber facilitado el acceso a clases virtuales y a materiales de apoyo educativos. Del mismo modo

factores como la deserción universitaria durante pandemia o el aplazamiento educativo mientras se controlaba la pandemia.

Para el año 2021(Tabla 7), segundo año de pandemia, se encuentra que el crecimiento en el puntaje evidenciado se ve estancado con disminuciones de 0,43% y 1,67% para Caldas y Quindío respectivamente. Esto no se evidencia para Risaralda que presenta un leve incremento de 0,64%. cambios se ven más consistentes para Caldas y Quindío que tienen una tendencia de descenso similar mientras Risaralda presenta un comportamiento diferencia de los otros dos departamentos de crecimiento leve en el puntaje de 0.64% (Grafica 3) esto puede suponer a como las instituciones y los estudiantes pudieron adaptarse al proceso de la virtualidad.

Estos resultados será necesario analizarlos en un mayor contexto, entendiendo que este grupo poblacional, en teoría, es el mejor preparado para afrontar la pandemia al tener acceso a clases virtuales sin necesidad de recurrir a ayudas externas o desplazarse a otras viviendas o territorios. También evidencia mayor facilidad a la hora de adaptarse a los cambios y, por último, pero no menos importante, hay una disminución en el número de personas presentando la prueba durante pandemia (Tabla 6), especialmente en el segundo año, que puede decantar aquellas personas que se sentían con menor preparación para la prueba como tal.

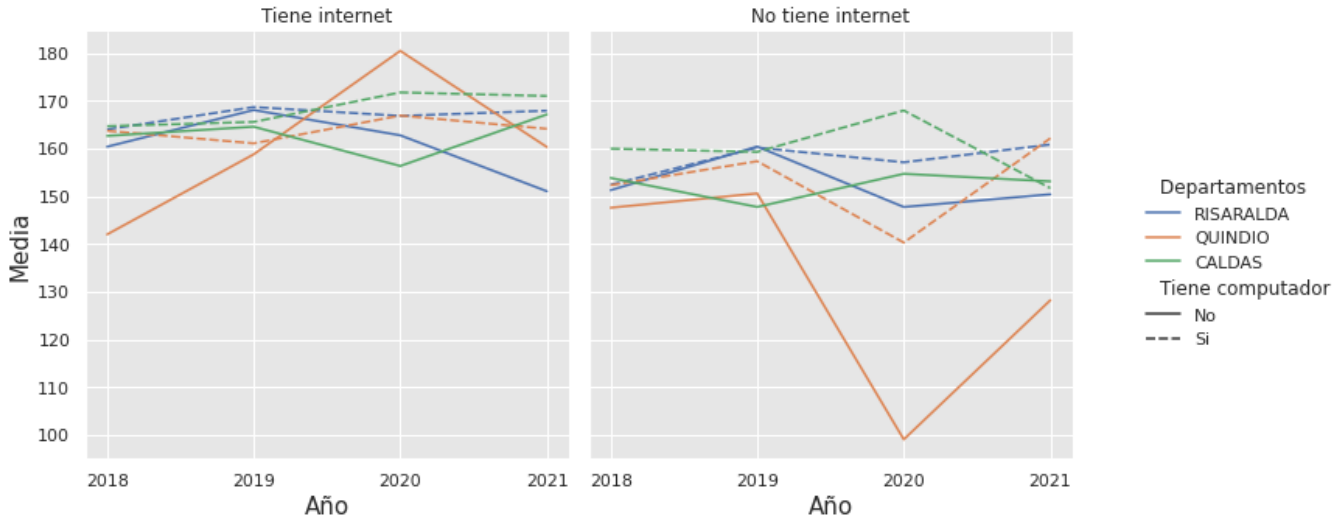
	Anio	2018		2019		2020		2021	
	FAMI_TIENECOMPUTADOR	No	Si	No	Si	No	Si	No	Si
FAMI_TIENEINTERNET	ESTU_INST_DEPARTAMENTO								
No	CALDAS	153,84	159,9839	147,7586	159,3091	154,7143	168	153,125	151,6957
	QUINDIO	147,6	152,3158	150,6	157,35	99	140,2857	128,2	162,1667
	RISARALDA	151,2727	152,4091	160,4286	160,2182	147,75	157,125	150,4167	160,85
Si	CALDAS	162,6667	164,7045	164,5882	165,5941	156,3529	171,7824	167,1667	171,0383
	QUINDIO	142	163,7325	158,8	161,1011	180,5	166,9286	160,3333	164,1485
	RISARALDA	160,4	164,0373	168,087	168,7116	162,7857	166,8691	151,0556	167,9424

Tabla 7 Promedio de las personas que presentaron las pruebas Saber Pro para la prueba de ingles

Aquellos que Tienen internet y no poseen computador

Este segundo punto tiene en consideración el acceso a internet sin la presencia de un computador. Siendo importante destacar que, el número de personas en esta categoría para todos los años es bastante reducido

para los tres departamentos. En Risaralda, el departamento con mayor incidencia de esta categoría, se presentan máximo 25 personas (para el año 2018) mientras que en el 2020 y 2021, años de pandemia, se presentan 14 y 18 personas en esta categoría (Tabla 6). Es posible inferir dos escenarios, las personas adquirieron equipos de cómputo para poder acceder a las clases o algunos sin equipo de cómputo prefirieron aplazar sus estudios para después de la pandemia.



Gráfica 3 Media de puntaje de ingles para cada departamento dependiendo de cierto factor atraves de los años

Caldas y Quindío a diferencia de Risaralda presentan aún menos estudiantes sin computador, con acceso a internet, teniendo solo 17 y 9 estudiantes como máximo (Tabla 6), sin mayores variaciones. En cuanto a puntajes, se encuentra que de forma general son entre dos y cinco puntos inferiores a quienes tienen tanto computador como acceso a internet. Sin embargo, se presenta un comportamiento atípico particular para el primer año de pandemia dónde los estudiantes con estas condiciones en el Quindío tuvieron 14 puntos superiores de promedio que aquellos con computador e internet (Tabla 7). Esto es atribuible al pequeño tamaño de la población y a valores particulares puntuales.

Omitiendo estás situaciones atípicas es posible determinar que la pandemia, como mínimo, mantuvo la diferencia entre quienes poseen y no poseen un equipo de cómputo, en algunos casos incrementando aún estos valores hasta por encima de los diez puntos. Es posible atribuir esta tendencia al acceso que tenían

estos estudiantes a los equipos de cómputo institucionales, esto es, las salas de sistemas de las diferentes universidades para realizar actividades, situación que cambió debido a la cuarentena.

Sin acceso a internet

El internet se ha vuelto un servicio por así decirlo básico dentro de los hogares y se realizan año tras año proyectos para llevar conectividad a zonas del país donde no llega esta, es por eso por lo que la población que expresa no tener una conexión a internet es mínima (Tabla 6).

Aquellos que no tienen internet y si poseen computador

Los estudiantes que manifestaron no contar con el servicio de internet, pero contar con el equipo de cómputo para el año 2018 a pesar de ser una población muy reducida se sitúan sobre los 150 puntos de promedio en los tres departamentos (Grafica 3). Sin embargo, si se realiza un análisis desde el 2018 hasta el 2021, el puntaje de este grupo poblacional (sin internet, pero con computador) con excepción de Caldas, tiene un crecimiento en cuanto a puntaje de 152 a 162 para Quindío y 152 a 160 para Risaralda (Grafica 3). Es importante destacar que este grupo poblacional tiene disminuciones en su participación. Como se ha explicado anteriormente, esto puede explicarse por una deserción generalizada o por los esfuerzos de las instituciones / gobierno, para incrementar la conectividad de los estudiantes.

Aquellos que no tienen internet y no poseen computador

Existe un grupo poblacional con aún más problemas de conectividad, esto es el de los estudiantes sin equipo de cómputo y sin conexión a internet. De este grupo es importante destacar su tamaño, siendo considerablemente pequeño y aún más en los años post pandemia con menos de cinco para el 2020 en los tres departamentos y menos de 12 para el 2021 (Tabla 6).

Con respecto a los años de pandemia es muy importante destacar que el número de personas sin equipo de cómputo disminuyo considerablemente, con reducciones para el 2020 en un 76% menos que los del año anterior para Caldas, 93% para Quindío y 86% para Risaralda (Tabla 6). Esto implicaría que por fuerza los estudiantes se vieron en la necesidad de adquirir equipos de cómputo para asistir a sus clases, afectando "positivamente" esta métrica.

Realizar análisis con esta información es complejo ya que, a pesar de la poca cantidad de estudiantes que presentaron el examen con estas condiciones, por ejemplo, para el año 2020 solo una persona en el Quindío manifestó no tener ni computador ni internet (Tabla 6), lo anterior hace difícil la generalización de la información Sin embargo, si es posible evidenciar que hay una diferencia en cuanto a puntaje de inglés entre quienes no tiene conexión a internet ni equipo de cómputo y aquellos que si cuentan con ambas. Esta diferencia se sostiene del período prepandemia al período en pandemia, incluso se incrementa un poco, siendo estás diferencias, en promedio superiores en más 10 puntos totales para cada año en los tres departamentos (Tabla 7).

Información sociodemográfica

ESTU_INST_DEPARTAMENTO	ESTU_AREARESIDE	2018	2019	2020	2021
CALDAS	Area Rural	5,50	4,51	6,67	7,14
	Cabecera Municipal	94,50	95,49	93,33	92,86
QUINDIO	Area Rural	5,92	7,23	8,59	9,83
	Cabecera Municipal	94,08	92,77	91,41	90,17
RISARALDA	Area Rural	7,12	7,14	11,54	9,75
	Cabecera Municipal	92,88	92,86	88,46	90,25

Tabla 8 Porcentaje de ubicación de la población estudiantil

Una de las características que se tienen en cuenta es también en donde reside el estudiante, si el estudiante reside en un área rural o si el estudiante reside dentro de la cabecera municipal. Esto debido a que todas las instituciones o la gran mayoría se ubican dentro de las ciudades principales de cada departamento, para este caso de estudio hay que tener en cuenta que más del 90 % de los estudiantes residen dentro de la cabecera municipal, dejando así a los estudiantes de áreas rurales como el 10 % de la población que presento el examen o menos (Tabla 8).

ÁREA RURAL

Para el área rural se han impulsado proyectos para llevar conectividad a aquellas zonas donde aún no llega el internet desde el decreto 555 de 2020 se determinan los servicios de comunicaciones como esenciales en busca de incrementar la conectividad que se estima en 1 de cada 6 hogares para la zona rural nacional

(Mintic, 2020). Siendo importante ya que la población rural, una que el Estado quiere tanto incentivar su acceso a educación superior como incentivar su permanencia en el área rural con conocimiento de alto nivel para mejorar las condiciones de la zona en la que reside. El número de estudiantes universitarios, no supera el 11% del total en ningún momento; sin embargo, es posible notar un ligero incremento en la participación rural para los años de pandemia, pasando de un promedio del 6% para 2018 – 2019, a un promedio de 9% para el 2020 – 2021(Tabla 8).

	Anio	2018		2019		2020		2021	
	FAMI_TIENEINTERNET	Si	No	Si	No	Si	No	Si	No
ESTU_AREARESIDE	ESTU_INST_DEPARTAMENTO								
Area Rural	CALDAS	71,43	28,57	80,00	20,00	81,63	18,37	86,96	13,04
	QUINDIO	93,75	6,25	83,33	16,67	95,24	4,76	95,45	4,55
	RISARALDA	94,83	5,17	77,19	22,81	81,44	18,56	87,50	12,50
Cabecera Municipal	CALDAS	89,44	10,56	90,08	9,92	95,71	4,29	95,62	4,38
	QUINDIO	90,94	9,06	89,49	10,51	97,39	2,61	95,10	4,90
	RISARALDA	93,12	6,88	90,20	9,80	96,49	3,51	95,89	4,11

Tabla 9 porcentaje de gente por zona que cuentan con internet

	Anio	2018		2019		2020		2021	
	FAMI_TIENEINTERNET	Si	No	Si	No	Si	No	Si	No
ESTU_AREARESIDE	ESTU_INST_DEPARTAMENTO								
Area Rural	CALDAS	30	12	28	7	40	9	40	6
	QUINDIO	15	1	20	4	20	1	21	1
	RISARALDA	55	3	44	13	79	18	56	8
Cabecera Municipal	CALDAS	618	73	690	76	647	29	567	26
	QUINDIO	231	23	264	31	224	6	194	10
	RISARALDA	690	51	663	72	715	26	583	25

Tabla 10 cantidad de estudiantes por zona

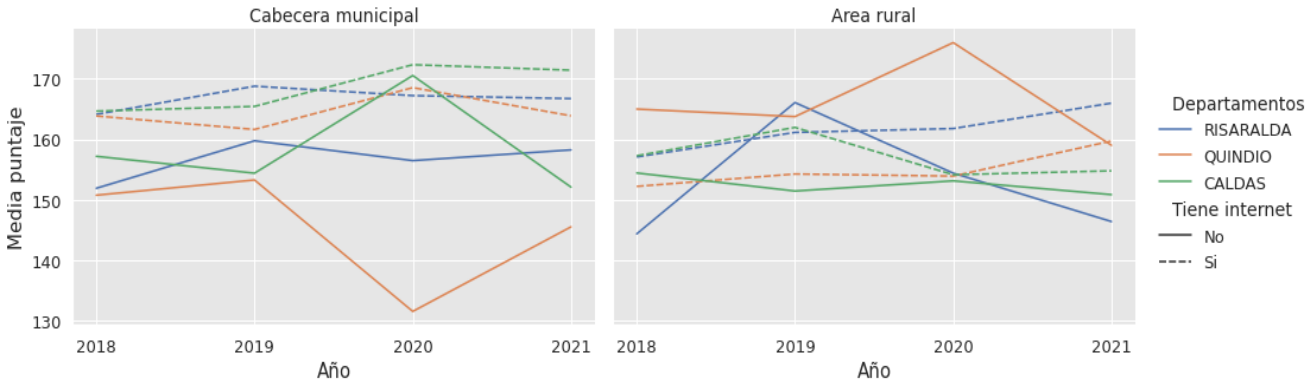
Este cambio es posible debido a los estudiantes que se desplazaron a la zona urbana para desarrollar sus estudios y en la virtualidad decidieron regresar a la zona rural. Desagregar la población rural entre quienes tienen acceso internet y los que no, muestra, en primer lugar, una buena conectividad tanto prepandemia como en pandemia, con cerca de un 95% - 93% (Tabla 9) de la población con acceso a internet y en segundo lugar presenta un problema de análisis para la comparación del desempeño al presentarse solo 1 estudiante incluso en algunos departamentos por año (Tabla 10).

	Anio	2018		2019		2020		2021	
	FAMI_TIENEINTERNET	No	Si	No	Si	No	Si	No	Si
ESTU_AREARESIDE	ESTU_INST_DEPARTAMENTO								
Area Rural	CALDAS	154,4167	157,2667	151,4286	161,9643	153,1111	154,125	150,8333	154,775
	QUINDIO	165	152,2	163,75	154,25	176	153,9	159	159,7143
	RISARALDA	144,3333	157,0727	166,0769	161,1364	154,3889	161,7722	146,375	165,9821
Cabecera Municipal	CALDAS	157,1781	164,6618	154,3816	165,4391	170,5517	172,3447	152,0769	171,4356
	QUINDIO	150,7391	163,8571	153,2581	161,6212	131,5	168,5446	145,5	163,8814
	RISARALDA	151,8824	164,1188	159,75	168,7873	156,4615	167,2266	158,24	166,7324

Tabla 11 Promedio de puntaje por zona

Hay patrones que se mantienen con los estudiantes de la cabecera municipal y los estudiantes del área rural que es posible analizar, por ejemplo, los estudiantes de las cabeceras municipales con conexión a internet presentan un promedio de puntajes superior a todos los demás en todos los años analizados antes y durante la pandemia (Grafica 4). Esto es consistente con la hipótesis que la presencia de conectividad contribuye a la obtención de mayores puntajes en las pruebas saber pro.

La cabecera municipal con conexión a internet obtuvo puntajes superiores a aquellos con conexión a internet en el área rural, esto puede estar vinculado a diferentes factores incluyendo el manejo de las cuarentenas, la disponibilidad de tiempo para el estudio por parte de los estudiantes, entre otros, exceptuando el departamento del Quindío en los años 2020 y 2021 (considerando que este puntaje representa un único estudiante) es posible encontrar que los estudiantes del área rural sin conexión a internet tienen un promedio de puntajes más bajo que sus pares del área urbana con conexión a internet durante los años de la pandemia.



Gráfica 4 media de puntajes de inglés para cada departamento dependiendo de su localización a través de los años

Finalmente, los estudiantes sin conexión a internet desagregados por zona rural y zona urbana durante la pandemia si mantuvieron las diferencias marcadas. Para los años prepandemia ambos grupos poblacionales presentaban promedios similares, aunque ligeramente inferiores para la zona rural, estas diferencias se mantuvieron y en algunos casos se incrementaron para los años de pandemia, especialmente para Caldas en el 2020 y para Risaralda en el 2021. Lo anterior en conjunto con la brecha que ya se presentaba entre aquellos con y sin conexión a internet deja a la población rural en el último lugar, tanto antes como después de la pandemia.

Distribución de las categorías de ingles

El Ministerio de Educación Nacional (M.E.N) dentro del marco de la política de educación nacional durante el año 2013 estableció como objetivo nacional la política de bilingüismo con el Proyecto de Fortalecimiento al Desarrollo de Competencias en Lenguas Extranjeras (2013) dentro del cual se busca el desarrollo de las competencias de las lenguas extranjeras para el fortalecimiento de la educación básica y superior en busca de "desarrollar competencias comunicativas dentro de las lenguas extranjeras especialmente dentro de la lengua inglesa aportando al capital humano en una economía en crecimiento y un mercado globalizado". Este proyecto plantea como objetivo nacional para los estudiantes de grado 11, un mínimo de 40% por encima del nivel B1, mientras para los estudiantes universitarios, fuera de las licenciaturas de inglés, se establece una meta de 20% en el nivel B2.

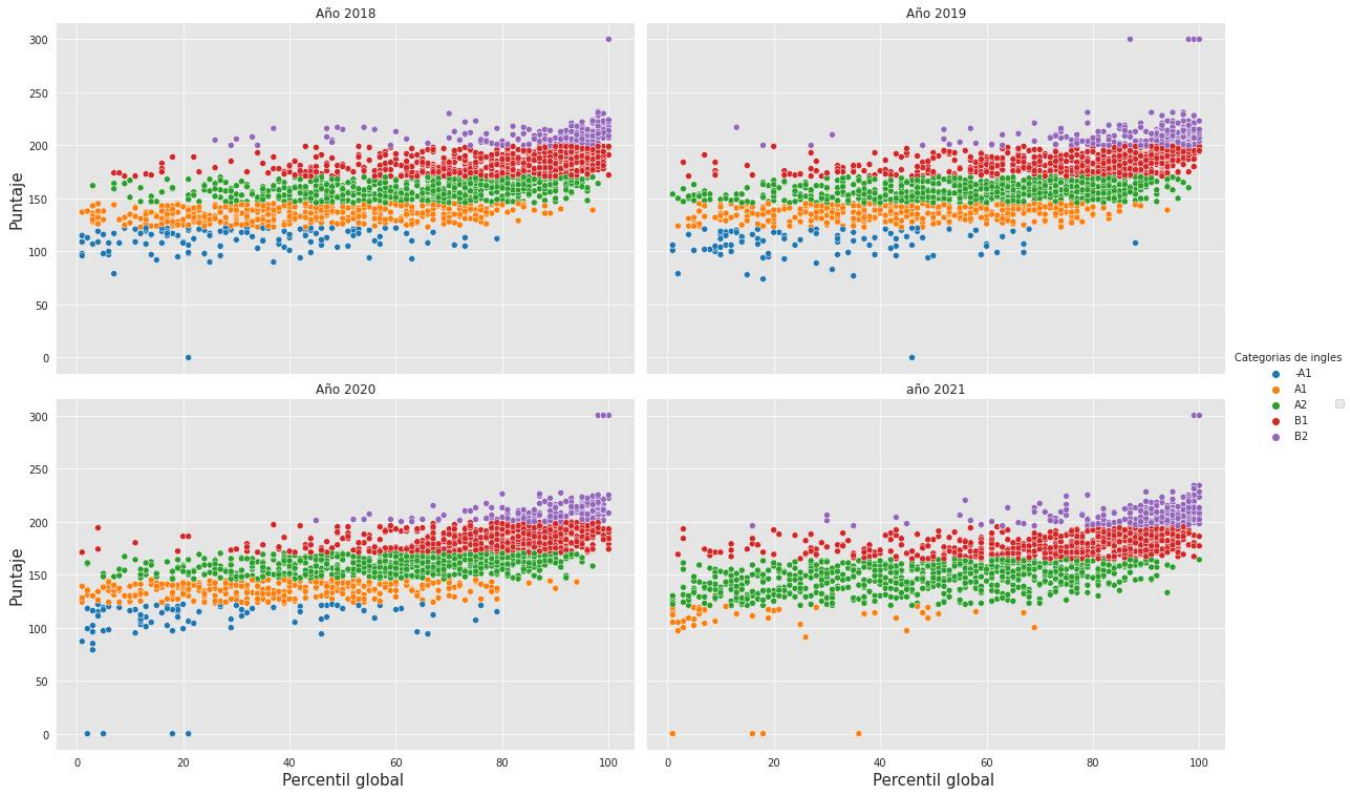
En el año 2014 se planteó el Programa Nacional de inglés 2015-2025 Colombia Very well (MEN, 2014) con la iniciativa de continuar con lo planteado en el 2013, en un establecimiento de metas específicas para el 2025 en las pruebas Saber Pro. Se establece el objetivo de tener al 30% de los estudiantes en el nivel B1 y un 25% en el nivel B2 para los estudiantes fuera de la licenciatura de inglés.

En este sentido, es importante analizar la evolución de los resultados antes y durante pandemia en cuanto al cumplimiento de la meta del programa nacional de inglés. A partir de la información en la tabla 12 es posible deducir que la meta para el nivel B2 no se ha cumplido en ningún momento del intervalo analizado. Sin embargo, para Caldas se ha presentado un incremento sostenido del porcentaje de estudiantes en el nivel B2 incluso en los años de pandemia. Sin embargo, los tres departamentos para el año 2021 han incrementado los estudiantes del nivel B2 para Quindío en el año antes de la pandemia presento una reducción leve en comparación a los años siguientes mientras que, para el caso de Risaralda, el primer año de pandemia hubo una reducción significativa en comparación a los resultados anteriores, este desempeño se ve corregido en el 2021 con un incremento por encima de la línea prepandemia.

ESTU_INST_DEPARTAMENTO	CALDAS				QUINDIO				RISARALDA			
Anio	2018	2019	2020	2021	2018	2019	2020	2021	2018	2019	2020	2021
MOD_INGLES_DESEM												
-A1	6,29	5,18	4,19	-	8,3	6,63	6,98	-	6,12	3,83	4,46	-
A1	22,26	19,16	15,43	2,12	19,03	23,19	17,44	4,27	21,88	15,33	17,72	3,6
A2	30,57	35,18	31,94	41,82	31,49	35,24	29,84	54,27	31,53	34,61	36,27	45,82
B1	31,45	28,43	34,24	37,73	33,91	28,92	36,43	27,78	30,47	34,25	32,63	37,61
B2	9,43	12,05	14,21	18,33	7,27	6,02	9,3	13,68	10	11,98	8,92	12,97

Tabla 12 porcentaje de gente en cada módulo de ingles

Con respecto a la meta de 30% de los estudiantes en el nivel B1 los resultados para Risaralda cumplen la meta, para Caldas este cumple igualmente a excepción del año 2019. Se encuentran por encima del objetivo, incluso es posible observar un incremento positivo que pasa de una línea cercana al 30% para el 2018 hasta un 37% para el 2021 para Caldas y Risaralda. El caso del Quindío es atípico con años por encima de la métrica y años por debajo de la métrica tanto antes como durante la pandemia.



Gráfica 5 Distribución de puntajes de ingles por clasificación de desempeño para cada año

Sin embargo, dentro de este análisis es importante destacar que la clasificación para el año 2021 suprime el nivel -A1. y los puntajes de ese nivel para años anteriores se agruparon en el nivel A1 como muestra la Gráfica 5. Así, hay una distribución de tipo normal de los resultados con un pico en el nivel A2, donde se cumple con la métrica nacional para el nivel B1, pero no para el nivel B2, indicando aun la necesidad de un crecimiento generalizado de los puntajes de los estudiantes para alcanzar ambas métricas antes de la fecha señalada y, a su vez, que la pandemia no ha tenido una incidencia negativa en la distribución de los estudiantes en los niveles establecidos.

Prueba de Hipótesis

Se realizo una prueba de hipótesis donde se quiere verificar observaciones sobre la pandemia en los resultados de los estudiantes, para este caso se utilizó una prueba de hipótesis con distribución t, debido a que tomaremos la población de antes de pandemia compuesta por los años 2018 y 2019 y la población en pandemia compuesta para los años 2020 y 2021 donde se analiza si la pandemia tuvo algún efecto en el desempeño del módulo de inglés de las pruebas Saber Pro.

Tenemos la formula general para los valores t

$$t = \frac{(x_a - x_b)}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

Para nuestro caso transformaremos la ecuación en la siguiente forma

$$t = \frac{(x_{Prepand} - x_{Pand})}{\sqrt{\frac{s_{Prepand}^2}{n_{Prepand}} + \frac{s_{Pand}^2}{n_{Pand}}}}$$

De los datos obtenemos los valores necesarios para poder obtener el valor de t, los cuales son

$N_{Prepand} = 3931$

$N_{Pand} = 3437$

$x_{Prepand} = 164.0661409310608$

$x_{Pand} = 167.13238289205702$

$\sigma_{Prepand} = 27.770115939203883$

$\sigma_{Pand} = 27.770115939203883$

Teniendo los valores procedemos a hallar el valor t:

$$t = -4.567482852448612$$

Obteniendo el valor t procedemos a plantear nuestra prueba de hipótesis

$H_0 \rightarrow El\ promedio\ de\ los\ puntajes\ de\ ingles\ de\ las\ pruebas\ saber\ pro\ en\ pandemia\ es\ igual\ que\ en\ pre\ pandemia$

$H_A \rightarrow El\ promedio\ de\ los\ puntajes\ de\ ingles\ de\ las\ pruebas\ saber\ pro\ en\ pandemia\ es\ mayor\ que\ en\ pre\ pandemia$

Estableceremos un valor Alpha α = 0.05 y los grados de libertad de 7366,

Realizando la prueba de hipótesis para hallar el valor p obtendríamos que

$$p = 0.000002508220911726662$$

donde ya teniendo el valor de p debemos tomar a consideración que

$$p \leq \alpha, \qquad entonces\ H_0\ es\ rechazado$$

Por lo cual teniendo en cuenta que el valor p es menor o igual a Alpha podemos rechazar la hipótesis nula y por ende concluimos entonces que el promedio de los puntajes del módulo de inglés en las pruebas Saber Pro en pandemia es mayor que en prepandemia

El dominio de una segunda lengua es un factor de relaciones sociales y productivas crucial, en las últimas décadas el inglés se ha considerado universal para las comunicaciones internacionales. Esto tienen importancia tanto para el mercado laboral como para el mundo académico, las pruebas internacionales como el TOELF o el IELTS se vuelven cada vez más centrales. Esta situación se incrementa cuando Latinoamérica se comienza a especializar en el ofrecimiento de servicios a nivel internacional, especialmente en las

empresas digitales, donde el mercado de empresas y los trabajadores encuentran en el teletrabajo una oportunidad.

Concluir sobre la incidencia que presentó la pandemia en las pruebas Saber Pro a partir de los resultados obtenidos.

Para comprender los efectos de la pandemia en el aprendizaje y/o dominio del inglés en los estudiantes universitarios. Fue necesario acceder a la información de las pruebas Saber Pro suministradas por el ICFES, es importante destacar el fácil acceso a la información gracias a la política de datos abiertos, el cual permite realizar los análisis y validar la evolución que presentan los datos en el tiempo.

Gracias a las series históricas que contiene la información sociodemográfica de los estudiantes se facilitó la agrupación y segmentación correspondiente de estos. Con estos datos fue posible encontrar variaciones presentes entre los años 2018 al 2021 en el núcleo académico de las ingenierías para los departamentos de Caldas, Risaralda y Quindío.

Entre los principales hallazgos del análisis de la información se encuentran: primero los cambios en la participación, segundo los cambios en la composición demográfica, y tercero los cambios en los puntajes y la relación entre estos. En primer lugar, se evidencia que la pandemia afectó de manera negativa a los tres departamentos en cuanto al total de personas participando en las pruebas en general, esto se entiende como una relación directa al aplazamiento de los estudios o deserción debido a la pandemia. En Risaralda en el 2020 se identifica un leve crecimiento del 2.0 % y una disminución del 18.5% en el 2021, Caldas con una disminución del 11% en el 2020 y en el año 2021 del 20.7%, para el departamento del Quindío en el 2020 de tiene una disminución del 22.2 % y para el 2021 una caída del 9.3%.

En segundo lugar, se tienen los cambios en la composición demográfica, donde se encuentra un desplazamiento de residencia de la población del área urbana al área rural, Esto se explica desde la oportunidad o necesidad de los estudiantes regresar a sus territorios de origen para continuar los estudios de forma remota si esto era posible. Los componentes de esta situación incluyen la facilidad para estudiar de manera virtual, así como la capacidad de acceder a equipos de equipo de cómputo e internet desde el hogar. En Risaralda en el año 2020 los estudiantes del área urbana disminuyen un 2.1% y para el 2021 un 17.4%,

para el caso de Caldas en el año 2020 disminuye un 12.5% y en el 2021 un 11% y, por último, en el Quindío se observa una disminución del 24% en 2020 y del 2021 un 9.9% durante el período de pandemia.

Por ultimo los cambios en los puntajes y la relación entre estos se logra evidenciar en los resultados expuestos anteriormente como la diferencia entre los puntajes de aquellos con acceso a internet y acceso a dispositivos de cómputo, en los tres departamentos. En este sentido, la ausencia de uno o ambos componentes (equipo de cómputo / conectividad) se presenta como una clara desventaja para estos estudiantes donde poseer ambos elementos supone mejores resultados, en este sentido es necesario revisar las estrategias de enseñanza de una segunda lengua implementadas durante la virtualidad.

Tal revisión debe fundamentarse en el incremento en los resultados de inglés en la prueba Saber Pro; que superan los resultados previos al periodo de la pandemia. Sin embargo, las metas establecidas para alcanzar el nivel B2 no se han logrado; debido a esta situación es importante desarrollar estrategias que faciliten la adaptación utilizando aquello aprendido durante la transición a la enseñanza en virtualidad en tiempos de pandemia, incluyendo, pero no limitándose a garantizar el acceso de forma virtual, pero también a la adaptabilidad de los materiales educativos, más en un área como el inglés.

De igual forma, representativo de la diferencia que supuso para los estudiantes durante pandemia no tener acceso a sus clases en modalidad virtual marcando aún más una diferencia que ya se hacía presente previo a la pandemia. De esto es posible analizar que la pandemia representó dos impactos, uno positivo en la disminución en términos porcentuales de personas sin conectividad y uno negativo en la disminución de los promedios totales de aquellos sin acceso a conectividad

En este sentido es importante, revisar las estrategias implementadas por las universidades o el gobierno departamental de Caldas quien logró mantener e incluso incrementar la cantidad de estudiantes que obtuvieron una certificación del idioma B1 y B2 durante el primer año de pandemia. Así como, entender las estrategias de recuperación de Risaralda y Quindío que, lograron recuperarse luego de un año bastante

complicando. Ilustrando un período de adaptabilidad más lento, pero con resultados positivos en el largo plazo, dentro del marco de tiempo analizado.

Conclusiones

El entendimiento de la metodología CRISP y el seguimiento de sus fases son herramientas poderosas a la hora de realizar procedimientos complejos de datos, esta metodología es orientada a la minería de datos donde los fases a tomar dan una guía que se puede moldear en diferentes objetivos finales dependiendo de las necesidades, y que evoluciona hacia la ciencia de los datos, lo cual toma de las metodologías de minería y lo junta procesos que permitan el análisis más a profundidad y la interpretación de la información de una forma más clara

Es importante tomar en cuenta como las modalidades educativas híbridas, o con estrategias más orientadas a la virtualidad pueden incidir en los resultados obtenidos, poder estudiar más a fondo las instituciones educativas y el desempeño de los estudiantes que presentan estos modelos y como afrontaron este periodo.

Por otra parte, es oportuno para llegar a conclusiones más precisas estudiar tanto las políticas que tomaron cada uno de los departamentos en materia educativa, en algunos casos entrando en las decisiones de cada universidad que podrían haber facilitado el acceso a clases virtuales y a materiales de apoyo educativos. Del mismo modo factores como la deserción universitaria durante pandemia o el aplazamiento educativo mientras se controlaba la pandemia.

Es necesario tomar en cuenta el cumplimiento parcial de las metas propuestas sobre el bilingüismo en los profesionales alcanzando los niveles de B1 pero quedándose muy por detrás para el nivel B2 dan a entender una problemática a nivel del dominio del lenguaje donde B1 el estudiante puede desenvolverse en situaciones sencillas, también teniendo en cuenta el gran porcentaje de estudiantes que están por debajo de estas dos clasificaciones dejan dudas claras sobre el dominio de una segunda lengua en la región cafetera y así mismo el nivel actual que tiene Colombia frente a este hecho

Finalmente, se recomienda continuar realizando un análisis de las pruebas subsecuentes para analizar los efectos residuales de la pandemia en los años posteriores, así como la posibilidad de aproximarse a la

pregunta de investigación desde un área más reducida, incluyendo las universidades como tal, lo cual facilitaría evidenciar estrategias puntuales de educación virtual que hayan implementado las universidades y su éxito o fracaso.

Bibliografía

Alonso, J. C., Estrada, D., & Martinez, D. (2016). ¿Se cumplió la meta de bilingüismo en los programas de educación universitaria del sector software en Colombia? *Revista Educación en Ingeniería, 11*(22), 39-45.

Alonso-Cifuentes, J., Estrada-Nates, D. and Mueces-Bedón, B., 2019. Evaluación del nivel de inglés en los programas de enfermería en Colombia: 2011-2016. *Revista Colombiana de Enfermería*, 18(2), p.e009

Alonso Cifuentes, J. C., Estrada Nates, D., & Mueces Bedón, B. V. (2018). Nivel de inglés en los programas de Administración de Empresas en Colombia: la meta está lejos. Estudios gerenciales, 445–456.

Alonso, J. C., Estrada, D., Mueces, B. V., & Sandoval-Escobar, M. (2018). Análisis de las competencias en segundo idioma en los programas de psicología colombianos. Suma Psicológica, 25(2).

Alonso Cifuentes, J. C., Estrada Nates, D., & Martínez Quintero, D. A. (2016). ¿Se cumplió la meta de bilingüismo en los programas de educación universitaria del sector software en Colombia? Revista Educación En Ingeniería, 11(22), 39-45

Azevedo, A., & Santos, M. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *Proceedings of IADIS European Conference on Data Mining*, (págs. 182-185). Amsterdam.

Bancroft, W. J. (1999). Suggestopedia and language acquisition: variations on a theme. Taylor & Francis

Beatty, K. (2013). Teaching & researching: Computer-assisted language learning.

Benati, A. (2018). Grammar-Translation Method. The TESOL Encyclopedia of English Language Teaching. https://doi.org/10.1002/9781118784235.eelt0153

Bermúdez, J. & Fandiño, Y. (2016). Bilingualism: definitions, perspectives, and challenges. In Ruta Maestra: Perspectives on English Language Teaching. Santillana.

Bermúdez Jiménez, J., Fandiño Parra, Y. & Ramírez Valencia, A. (2014). Percepciones de directivos y docentes de instituciones educativas distritales sobre la implementación del Programa Bogotá Bilingüe. Voces y Silencios. Revista Latinoamericana de Educación, 5(2), 135-171.

Bloomfield, L. (1933). Language. New York: Holt, Rinehart & Winston.

Cabeza, L., Lombana, J., & Castrillón, J. (2020). Factores externos en el desempeño de pruebas genéricas de Estado (SaberPro) de inglés en estudiantes de administración y afines en Colombia. Revista Iberoamericana de educación superior, 11(30).

Chadha, A. (2018). Efficent clustering algorithms in Educational Data Mining. En A. Malheiro, F. Ribeiro, G. Leal, J. Pocas, & O. Mealha, *Handbook of research on knowledge management for contemporary business environents* (págs. 279-312). IGI GLOBAL.

Chapman, P., Clinton, J., Kerber, R., Khazaba, T., Reinartz, T., Shearer, C., & R., W. (2000). CRISP-DM 1.0 step by step data mining guide. *CRISP-DM consortium*.

Cuenca, A. (2016). Desigualdad de oportunidades en Colombia: impacto del origen social sobre el desempeño académico y los ingresos de graduados universitarios. *Estudios pedagógicos* (2), 69-93

Fansury, A. H., Januarty, R., & Ali Wira Rahman, S. (2020). Digital Content for Millennial Generations: Teaching the English Foreign Language Learner on COVID-19 Pandemic. Journal of Southwest Jiaotong University, 55(3), 2-10 https://bit.ly/3mWR5OQ

Frawley, W., Piatetsky-Shapiro, G. y Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. AI Magazine, 13(3), 58

Gallegos Ibarra, I. P. (2022). COVID-19 Pandemic's Impact on English Teaching.

Kaur, R., & S., S. (2016). A survey of data mining and social network analysis based anormaly detection techniques. *Egyptian informatics journal, 17*(2), 199-216.

Khasanah, A. U. & Harwati. (2019). Educational Data Mining Techniques Approach to Predict Student's Performance. *International Journal of Information and Education Technology*, *9*(2), 115-118. https://doi.org/10.18178/ijiet.2019.9.2.1184

Lam, A. 2001. "Bilingualism". In R. Carter & D. Nunan (eds.) The Cambridge Guide to Teaching English to Speakers of Other Languages. Cambridge: Cambridge University Press. 93-100.

Larose, D. y Larose, Ch. (2014). Discovering Knowledge in Data: An Introduction to Data Mining (2da. ed.). New Jersey: John Wiley & Sons.

López Naranjo, H. A., & Sellamen Garzón, A. (2019). Determinantes del nivel de inglés en la Educación Superior en Colombia. Revista CIFE Lecturas de Economía Social, 21(34), 69–91.

Mejía-Mejía, S. (2016). ¿Vamos hacía una Colombia bilingüe? Análisis de la brecha académica entre el sector público y privado en la educación del inglés. *Educ. Educ., 19*(2), 223-237.

Merchán, J., *et al.* (2021). *La importancia del bilingüismo para los futuros egresados de ingeniería en la Universidad Ean y la Universidad Nacional* [Documento de trabajo, Universidad EAN]. Recuperado de: http://hdl.handle.net/10882/10998.

Ministerio de Educación Nacional. (2015). Programa Nacional de Inglés 2015-2025 "Colombia very well". Bogotá D.C. Obtenido de https://www.mineducacion.gov.co/1759/articles-343837_Programa_Nacional_Ingles.pdf

MINTIC Colombia (2020.). *El nuevo plan del Gobierno para conectar a las zonas rurales*. https://mintic.gov.co/portal/inicio/Sala-de-prensa/MinTIC-en-los-medios/135808:El-nuevo-plan-del-Gobierno-para-conectar-a-las-zonas-ruraleslas+zonas+rurales

Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2012). Estudio comparativo de metodologías para minería de datos. In *XIII Workshop de Investigadores en Ciencias de la Computación*.

Mohamad, S., & Tasir, Z. (2013). Educational data mining: A review. *Procedia - Social behavioral sciences, 97*, 320-324.

Oviedo, A., & Jiménez, G. (2018). Estudio sobre estilos de aprendizaje mediante minería de datos como apoyo a la gestión académica en instituciones educativas. *RISTI - Revista Iberica de Sistemas y Tecnologías de Información, 29*, 1-13.

Oviedo, A., & Jiménez, J. (2019). Minería de datos educativos: análisis del desempeño de estudiantes de ingeniería en las pruebas Saber Pro. *Revista politécnica, 15*(29), 128-140.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications, 41*(4), 1432-1462.

Rengifo, J. S., Sánchez, C., Delgado, C., Solarte, C., Vidal, F., & Timarán, R. (2020). Analítica de datos aplicada al contexto universitario. Caso de estudio: pruebas Saber Pro. *Cuaderno Activa, 12*, 13-19.

Reyes, J. (2011). Las estrategias discursivas de grupos sociales en la Universidad Nacional de Colombia en relacion con la Escritura académica en situaciones de bilingilismo e Interculturalidad (tesis). Universidad Nacional de Colombia, Bogota, Colombia.

Richards, J. C. (2005). Communicative language teaching today. Cambridge University Press. Tafazoli, D., Huertas Abril, C. A., & Gómez Parra, M. E. (2019). Technology-based review on Computer-Assisted Language Learning: A chronological perspective. Pixel-Bit: Revista de Medios y Educación, 54, 29-43. https://doi.org/10.12795/pixelbit.2019.i54.02

Segura, A., Trujillo, F., Álvarez, D., Postigo, A. Y., Fernández, M., Montes, R., & Trujillo, J. (2020). Aprender y enseñar en tiempos de pandemia. Catarata

Schröer, C., Kruse, F. & Marx, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526-534.

Trcka, N., Pechenizkiy, M., & van der Aalst, W. (2011). Process mining for educational data. En C. Romero, S. Ventura, M. Pechenizkiy, & R. Baker, *Handbook of Education Data Mining* (págs. 123-142). Florida: Taylor and Francis Group.

Utami, T. P. (2020). An analysis of teachers' strategies on English e-learning classes during COVID-19 pandemic (a qualitative research at MTs Sudirman Getasan in the Academic Year 2019/2020). https://bit.ly/3mWpcWW

Vellido, A., Castro, F., & Nebot, Á. (2011). Clustering educational data. En C. Romero, S. Ventura, M. Pechenizkiy, & R. Baker, *Handbook of Educational Data Mining* (págs. 75-92). Florida: Taylor and Francis Group.

Zorčič, S. (2020). Dimensions of Remote Learning during the Covid-19 Pandemic in Minority Language Schools (The Case of Austrian Carinthia). *Razprave in Gradivo: Revija za Narodnostna Vprasanja*, (85), 223-252.