Factores que inciden en el desempeño de los módulos en competencias genéricas Saber Pro entre los años 2012 y 2019: modelo de minería de datos para la gestión curricular. Informe Final

David Alberto García Arango
Oscar Andrés Cuéllar Rojas
Cesar Felipe Henao Villa

Introducción

La comprensión de la dinámica de las instituciones de educación superior, y en particular de los diferentes programas formativos, es de carácter complejo y depende de una gran cantidad de factores, relaciones, variables, componentes, procesos y sujetos, que en forma sistémica interactúan para el cumplimiento de propósitos y fines plasmados en el proyecto educativo institucional.

Una de las aristas de este entramado se relaciona directamente con los resultados de las pruebas Saber Pro y la forma en que éstos reflejan el valor agregado aportado al egresado desde cada uno de los programas académicos. Un componente de este valor agregado se estudia mediante la comparación de los resultados de las pruebas Saber 11 y las pruebas Saber Pro, específicamente en las competencias genéricas de un contexto o región delimitado denominado conjunto INBC. Este conjunto está conformado por las 15 instituciones con resultados promedio de sus pruebas Saber 11 más cercanos entre sí y con 0,3 desviaciones estándar de distancia para un núcleo básico de conocimiento determinado.

El concepto de valor agregado propuesto por el ICFES cobra mayor fuerza y sentido en el marco del decreto 1330 de 2019 en el cual se establece la necesidad de articular los resultados de aprendizaje a los niveles micro, meso y macro curriculares, como una forma de identificar el aporte de los procesos sustantivos de la institución y de los programas en el marco de las políticas de aseguramiento de la calidad (Ministerio de Educación Nacional, 2019).

Con este panorama y considerando las opciones de acceso a las bases de datos de resultados de las pruebas que provee el ICFES, se presenta en el presente proyecto el estudio del cruce de resultados entre las pruebas Saber 11 y las Saber Pro; esto, con el objetivo de identificar correlaciones en variables sociodemográficas y dimensiones de competencias genéricas para programas académicos de ingeniería. La metodología, con enfoque mixto, se basa en la consolidación de 1’048.575 registros que abarcan los resultados de las pruebas saber Pro para los módulos de razonamiento cuantitativo, lectura crítica, competencia ciudadana, inglés y comunicación escrita; se toma como referencia los resultados de las pruebas Saber 11 y Saber Pro desde 2012-1 hasta 2019-2.

Con base en el contexto anterior, y considerando las oportunidades en cuanto a técnicas de minería de datos; en la investigación se busca generar escenarios predictivos de desempeño en las pruebas Saber Pro en los ámbitos institucional, departamental y regional, que aporten al desarrollo de políticas en materia de gestión curricular en Instituciones de Educación Superior - IES. Para ello, se propone en un primer momento la identificación de factores predictores de los resultados de las pruebas Saber Pro para los módulos genéricos entre los años 2012 y 2019. Posteriormente, se pretende analizar las relaciones o correlaciones existentes entre variables cualitativas o cuantitativas incidentes, como marco para la generación de un modelo predictivo de resultados de las pruebas Saber Pro entre los años 2012 y 2019. Finalmente, se espera llevar a cabo la evaluación de los resultados de aplicación de técnicas basadas en minería de datos para los resultados de las

pruebas Saber Pro, como apoyo en el desarrollo de políticas en materia de gestión curricular para las Instituciones de Educación Superior- IES en Colombia.

Lo anterior, busca aportar a la comprensión de las siguientes preguntas de investigación:

¿Cómo reinterpretar los factores predictores que intervienen en el desempeño de los módulos de competencias genéricas de las pruebas Saber Pro entre los años 2012 y 2019? y, ¿Cómo aplicar estos análisis en la gestión curricular de Instituciones de Educación Superior - IES?

Como resultado de la investigación se identifica que si bien es cierto las pruebas Saber 11 son predictores de los resultados de las pruebas Saber Pro, también existen variables predictores socioeconómicas que pueden ser motivo de estudio para el desarrollo de políticas en materia curricular tanto para las Instituciones de Educación Superior como para el Ministerio de Educación, el presente estudio identificó el costo de la matrícula y el hecho de contar con servicio de internet como aspectos que se corresponden a los diferentes niveles de desempeño en las pruebas Saber Pro.

Revisión de literatura y estado del arte

A continuación, se presentan algunas conceptualizaciones que apoyan la obtención de los objetivos propuestos en el proyecto en su nivel exploratorio.

Competencias genéricas y conocimiento

El debate sobre el conocimiento en relación con los nuevos retos que impone la sociedad actual, ha estado activo durante los últimos años en cuanto a políticas de educación superior. Este se encuentra enfocado en la mejora de la calidad educativa, considerando la educación superior como el subsistema al que la sociedad encarga la tarea de formar capital profesional, social y humano; de ampliar capacidades y opciones; de educar a las personas para que sean parte de una ciudadanía libre y crítica, y de desarrollar conocimiento experto (Pedraza, 2020).

En el ámbito laboral, se exige a los graduados una amplia gama de competencias y habilidades de resolución de problemas, cooperación e interacción, además de habilidades profesionales que se encuentran en constante actualización (Torres, Moreno & Rivas, 2021). Por tanto, la iniciativa de la Organización para la Cooperación y el Desarrollo Económicos - OCDE para determinar la viabilidad de una evaluación de los resultados de aprendizaje en educación superior, plantea que las políticas deben estar orientadas a mejorar la enseñanza, aprendizaje y programas de estudio; y a su vez, mejorar las habilidades docentes, el liderazgo y el compromiso (García, 2020).

Uno de los principales objetivos del proceso de aprendizaje es desarrollar la experiencia de los estudiantes en su campo, la cual se basa tanto en conocimientos y habilidades del campo específico de formación, como en competencias genéricas. Hay muchos tipos de habilidades genéricas, pero la educación superior generalmente se enfoca en las habilidades cognitivas, como la capacidad de pensar críticamente y argumentar, para inferir analíticas y tomar decisiones informadas (Feng & Wei, 2019).

En el marco de lo anterior, la taxonomía de Bloom se utiliza para identificar las habilidades genéricas mencionadas, orientadas en analizar, sintetizar y evaluar datos (Irvine, 2017). Así mismo, estudios previos han demostrado que los estudiantes enfrentan desafíos, por ejemplo, en

argumentación, evaluación de conocimientos y liderazgo en la toma de decisiones (Hyytinen, Toom, y Shavelson, 2019) y, además, existen factores socioeconómicos que inciden en ese desempeño. Las competencias genéricas están compuestas por varios aspectos como: (1) razonamiento analítico y evaluación, (2) resolución de problemas, (3) escritura argumentativa y (4) dominio del idioma. Estas competencias son clave en la formación integral de los nuevos profesionales (Martínez & González, 2018), por lo cual es importante considerar el papel de los factores socioculturales en la formación de éstas.

En los últimos años las competencias genéricas se incluyen tanto en los planes de estudios universitarios como en el debate sobre políticas de educación superior (Villamil, 2019). Se observa, entonces, que la vida laboral está cambiando rápidamente y que la gestión de competencias sectoriales no es suficiente para la educación superior. La experiencia, la resolución de problemas y la interacción se han convertido, así, en habilidades clave para la educación superior. Hoy en día, la tarea de las futuras investigaciones en estudios universitarios no sólo se orienta en profundizar sobre la experiencia en la construcción y aprendizaje de conocimientos sectoriales, incluido el aprendizaje de competencias genéricas; sino que, además, debe enfocarse en que los graduados obtengan nuevos aprendizajes en un campo laboral en constante cambio, proceso en el cual las competencias genéricas son fundamentales para el aprendizaje permanente (Karlgren, et al., 2020).

Por tanto, es necesario reconocer que las competencias genéricas son habilidades que permiten beneficiarse de la experiencia y los conocimientos técnicos sectoriales en su campo, tanto durante los estudios como posteriormente en la vida laboral. Por lo anterior, están ligadas indefectiblemente a diferentes factores sociales propios del ser humano. (Lindblom-Ylänne, Parpala y Postareff, 2019). Por ejemplo, las competencias orientadas a la escritura permiten al ser humano hacer visibles a otros sus pensamientos y conclusiones; las habilidades de razonamiento y argumentación, por otro lado, ayudan a defender los propios puntos de vista. Es así, como se ha descubierto que las habilidades genéricas permiten el éxito académico, el progreso de los estudios y el aprendizaje basado en la comprensión (Zapatera, 2021).

En este contexto, el presente proyecto busca identificar los factores sociales, económicos y culturales que inciden en el desempeño de los estudiantes, utilizando como eje transversal la aplicación de modelos de minería de datos como una estrategia integradora de las variables que allí confluyen.

El concepto de valor agregado y calidad en las pruebas estandarizadas: identificación de variables asociadas

El ICFES define el valor agregado como

> cuánto aporta una institución a las competencias de sus estudiantes. Cuando un alumno entra a un centro educativo llega con unas habilidades previas, por lo que sus capacidades al terminar su ciclo académico no se deben solo a lo que aprendió por su paso en él, sino también a lo que sabían antes de entrar al mismo. Por este motivo surgen los estudios de VA, los cuales intentan aislar lo aprendido en una institución de las condiciones iniciales, para poder medir de una forma más precisa la calidad en la formación académica. (ICFES, 2021, p. 1)

Martínez, Gaviria y Castro (2008), definen los modelos de valor agregado (VA) como especificaciones estadísticas a partir de las cuales se estima la efectividad educativa, haciendo énfasis en el progreso de los estudiantes a través del tiempo. Por tanto, el VA se puede entender como la contribución de las instituciones al progreso neto de los estudiantes, enfocada en objetivos de aprendizaje establecidos, una vez se controla la influencia de factores ajenos a la institución que pueden contribuir a este avance.

Según el ICFES (2021) "los exámenes Saber 11° y Saber Pro ofrecen mediciones comparables sobre el nivel de competencias con el que los estudiantes inician y finalizan el ciclo de educación postsecundaria, lo cual permite evidenciar el valor agregado del sistema de educación superior en Colombia" (p. 4).

Ahora bien, en el marco de la propuesta del Ministerio de Educación Nacional - MEN para la calidad de los programas formativos, es importante considerar que "los resultados de aprendizaje son concebidos como las declaraciones expresas de lo que se espera que un estudiante conozca y demuestre en el momento de completar su programa académico", y "se espera que los resultados de aprendizaje estén alineados con el perfil de egreso planteado por la institución y por el programa específico". (Ministerio de Educación Nacional, 2019)

Una de las características principales del decreto 1330 de 2019, el cual regula aspectos de acreditación en alta calidad, presenta el concepto de resultados de aprendizaje como un factor importante en el proceso asociado a la cultura de la autoevaluación de programas e instituciones. En este decreto se definen los resultados de aprendizaje como "las declaraciones expresas de lo que se espera que un estudiante conozca y demuestre en el momento de completar su programa académico" (Ministerio de Educación Nacional, 2019, p.4). Es en este sentido que muchas instituciones de educación superior, generan mecanismos para evaluar el aporte de la formación universitaria en estos resultados; identificando niveles de acercamiento de los resultados de aprendizaje del egresado al perfil deseable declarado en el programa. Así pues, la identificación de los niveles de avance se hace complejo de comprender teniendo en cuenta que los egresados, en la mayoría de los casos, se alejan de la institución universitaria, y dejan de aportar información de gran importancia para evaluar su desempeño.

En lo referente al valor agregado de la formación universitaria, el texto de Bogoya, Bogoya, & Peñuela (2017) ofrece una comparación entre modelos estadísticos, para identificar cuál de ellos es más preciso identificando relaciones o comportamientos de relación entre variables; indicando modelos predictivos sobre el nivel de aporte de las instituciones; e identificando aspectos de desigualdad en vertientes socioeconómicas. En este mismo contexto, la investigación de Melo, Ramos & Hernández (2014) identifica factores de eficiencia a través de técnicas de frontera estocástica en los resultados de las pruebas Saber Pro. Como resultado se determina que los factores socioeconómicos de los estudiantes limitan la influencia de los factores asociados a los puntajes.

Es fundamental resaltar que en el contexto nacional, el libro de la OCDE (2012) en el cual se analizan los logros de la última década y los desafíos en el intento de ofrecer un sistema educativo de clase mundial; establecen que la prueba Saber 11 del ICFES, a diferencia de lo que suele pensarse, no es suficientemente confiable para tener una idea del desempeño individual de los estudiantes. Por esto, las instituciones de educación superior tienen que reconocer el gran valor potencial de las pruebas Saber Pro (OCDE, 2012, p.59).

En este orden de ideas, es pertinente considerar los esfuerzos de la OCDE en la integración de los resultados de las pruebas por países, para identificar relaciones a nivel socioeconómico mediante el programa de evaluación de los resultados de aprendizaje para la educación superior (OCDE, 2012). Basándose así, en el AHELO -por sus siglas en inglés- de países como Emiratos Árabes, Australia, Bélgica, Canadá, Colombia, Egipto, Finlandia, Italia, Japón, Corea, Kuwait, México, Holanda, Noruega, Rusia, Eslovaquia y Estados Unidos. Para el caso específico de Colombia, se recalca la importancia del análisis relacional entre las pruebas Saber 11, las pruebas Saber Pro y los respectivos datos socioeconómicos, como una forma de producir medidas de valor agregado en instituciones de educación superior, y consecuentemente, comprender mejor los factores predictivos del desempeño para introducir mejoras en materia de gestión curricular, lo cual es el aspecto central del presente estudio.

Agudelo, Figueroa y Vásquez (2019), abordan los conceptos de calidad educativa para las pruebas estandarizadas en Colombia, en las modalidades presencial tradicional y virtual-distancia universitaria. Analizan, además, los datos de desempeño de una universidad del área metropolitana de Medellín con modalidad virtual-distancia, cuyos resultados fueron bajos durante los años 2015 y 2016. A través de un enfoque de corte positivista encuentran un grupo de variables asociadas al desempeño de los estudiantes en la prueba Saber Pro. Después de analizar los datos, los autores coinciden en que **los resultados Saber 11 son un determinante en la obtención de resultados Saber Pro por encima de la media,** ya que al separar los resultados de los estudiantes por modalidad de programa a distancia y presencial, los autores encontraron que la modalidad del programa incide en los resultados; al igual que la edad del estudiante y el tiempo transcurrido entre el bachillerato y el ingreso a la universidad.

Cifuentes, Chacón y Fonseca (2019), analizan las diferencias en los resultados de los exámenes de calidad de educación superior Saber Pro de los estudiantes de la Licenciatura en Educación Básica con énfasis en Matemáticas, Humanidades y Lengua Castellana. De modo que realizan un análisis descriptivo a partir de datos cuantitativos entregados por la plataforma ICFES; y encuentran diferencias entre los resultados de los estudiantes del programa según los módulos establecidos. Se hace necesario, entonces, reformular a través de estrategias pedagógicas, metodológicas y didácticas, la formación de los Licenciados en Educación Básica, con miras a fortalecer las competencias relacionadas con la enseñanza, evaluación y formación propias de los licenciados. Los desempeños en estos módulos específicos se encontraron por debajo de los promedios del grupo de referencia -Educación- y promedio nacional, de manera que el programa profesional en el campo de la educación puede mejorar en el proceso de medición por parte de las pruebas estandarizadas, para efectos de relación entre calidad y medición.

Desde otra perspectiva, se considera que el entorno familiar y las condiciones económicas y sociales propias de las familias influyen en el desempeño académico de los estudiantes, y, por ende, en los resultados de las pruebas académicas. Por su parte, García-González y Skrita (2019) predicen el desempeño académico de los estudiantes que presentaron el examen de Estado de 2016 para acceder a la educación superior (Saber 11) a partir de las observaciones y características familiares propias de los estudiantes. Los resultados muestran que las variables familiares que mejor predicen los resultados académicos son: **nivel educativo de la madre, estrato socioeconómico de la vivienda, número de libros, nivel educativo del padre y poseer computador en la vivienda.**

En cuanto a competencias, Niebles, Martínez-Bustos y Niebles-Núñez (2019), analizaron las competencias matemáticas como factor de éxito en la realización de la Prueba Saber Pro de

Universidades de Barranquilla, Colombia, en este análisis confirmaron que las competencias matemáticas, así como sus dimensiones, competencias genéricas y razonamiento cuantitativo, "casi siempre" son promovidas en los procesos de enseñanza. Este es un rango considerado como favorable.

En Colombia, las pruebas de Estado Saber Pro han sido diseñadas para apoyar la evaluación y el mejoramiento de la educación superior en el país. A través de la aplicación de la metodología de minería de datos CRISP-DM, se realizó un estudio de los resultados obtenidos en las pruebas Saber Pro de estudiantes de ingeniería en Antioquia (Colombia). A partir de 108 variables académicas, económicas y sociodemográficas se realizaron 3 modelos analíticos: 1) agrupación de los tipos de estudiantes, 2) selección de los factores que más influyen en el desempeño de las pruebas, y 3) predicción del desempeño en las pruebas a partir de las variables seleccionadas. Como resultado se encontró que algunas de las variables más influyentes para el resultado de las pruebas son: número de personas a cargo, método de enseñanza, si el hogar es permanente, carácter académico de la institución y facilidades económicas como tener horno, microondas, gas y motocicleta (Oviedo y Jiménez, 2019).

Han sido diversos los esfuerzos realizados para identificar factores o variables que inciden en el desempeño de los estudiantes de programas de pregrado en las pruebas Saber Pro. Esto, debido a que las pruebas se constituyen como un sensor que da cuenta de una de las muchas dimensiones de la complejidad asociada a los resultados de aprendizaje y los niveles de desarrollo de competencias que se asocian a los diferentes programas formativos.

En este escenario, Vásquez (2018) concluye que en materia de investigación se han realizado múltiples estudios que identifican diferentes factores que inciden en los resultados de las pruebas Saber, además, identifican que los factores que intentan establecer relaciones entre las pruebas Saber 11° y Saber Pro, no son concluyentes respecto a la predicción de los resultados de las segundas, tomando como variable independiente los resultados de las primeras. Los resultados del ejercicio empírico reafirman la importancia de las variables socioeconómicas en el logro académico de los estudiantes de educación superior. Ello sugiere que, aunque muchas instituciones educativas tienen margen para mejorar sus niveles de eficiencia, están restringidas por la influencia de los factores del entorno de sus estudiantes. De modo que para lograr un mejoramiento de los resultados académicos, las medidas de política del Estado y las estrategias de las instituciones deben tomar en cuenta; no sólo los criterios en la contratación de docentes, sino también la definición de incentivos para la investigación, los aspectos administrativos y financieros, y los mecanismos que permitan ayudar a contrarrestar el impacto negativo derivado de las condiciones socioeconómicas de los estudiantes y de otros factores ambientales (Melo, Ramos & Hernández., 2014).

Una medida de análisis de desempeño de graduados de los programas de formación universitaria en Colombia, es el resultado obtenido en las pruebas Saber Pro. Pues a menudo el análisis de los bajos resultados respecto a la media nacional en los módulos genéricos y específicos, es justificado en relación con la escasa o nula preparación de los estudiantes desde el bachillerato o las condiciones socioeconómicas, generando una multiplicidad de hipótesis al respecto.

En este contexto, una dimensión que aporta al conocimiento del nivel diferencial de avance en el marco de los resultados de aprendizaje, es el resultado de las pruebas Saber Pro tanto en los módulos de competencias genéricas como en los de específicas, además del establecimiento de patrones de relación con las pruebas Saber 11. El análisis de estos resultados debe ir más allá de

una interpretación porcentual y variacional comparativa; un análisis más complejo implica la consideración de un sinnúmero de variables analizadas en un modelo integrador, que antes de identificar como culpables los grados o procesos previos, permiten comprender interrelaciones que van más allá del promedio de las puntuaciones en un saber disciplinar para la comprensión del rol de las instituciones de educación superior en los contextos y realidades locales, para direccionar y autorregular procesos orientados a la calidad de los programas. En concordancia con lo anterior, esta investigación aportará bases para comprender mejor las dinámicas de la gestión curricular.

Metodología y datos

El desarrollo metodológico del proyecto se enmarca en un nivel de razonamiento abductivo con enfoque cuantitativo, pero con triangulación cualitativa, llevando a la consideración de un enfoque mixto. En relación con el proceso de razonamiento abductivo, se siguen los referentes propuestos por Samaja (2012), según los cuales la extracción del caso de estudio se genera al momento de relacionar los resultados obtenidos con la teoría existente. Este enfoque, cuyo nivel paradigmático está centrado en el pragmatismo, permite la configuración y articulación de métodos que posibiliten la integración de los resultados al contexto teórico. Por consiguiente, se concibe la importancia de analizar correlaciones y relaciones entre variables asociadas al desempeño en las pruebas Saber Pro y los resultados de aprendizaje, mediante la construcción, aplicación, validación y análisis de un modelo o conjunto de modelos predictivos, los cuales en sí mismos se conciben como "el proceso de aplicar un modelo estadístico o algoritmo de minería de datos con el propósito de predecir nuevas observaciones o corroborar las existentes". (Shmueli, 2010).

La minería de datos (MD) busca dar un sentido a los grandes volúmenes de información que puede ser almacenada; a través de ella es posible encontrar irregularidades en los datos y predecir resultados (Riquelme, Ruiz y Gilbert, 2006). El desarrollo de los modelos predictivos consistirá entonces, en la aplicación de un algoritmo de minería de datos, cuya aplicación está enmarcada en las fases de extracción de conocimiento como se evidencia en la Figura 1.

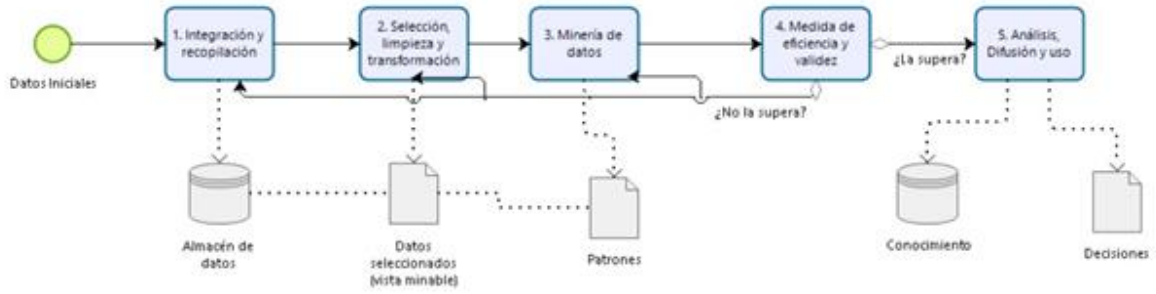


Figura 1. *Fases de extracción del conocimiento para la evaluación integral del proyecto.* Fuente: Construcción propia.

En el subproceso de minería de datos, los expertos sugieren el modelo de metodología Cross Industry Standard Process for Data Mining (CRISP-DM), el cual, si se adapta al fenómeno de estudio, presenta el ciclo que puede verse en la Figura 2. Con base en lo anterior, se tomarán las bases de datos de los resultados de las pruebas Saber Pro entre los años 2012 y 2019, y las respectivas pruebas asociadas Saber 11 utilizando como aspecto clave la base de datos denominada

"Llave_Saber11_2006-1_2019-2_SaberPRO_2012-1_2019-2.txt", la cual contiene 1.428.265 datos y permite la articulación de otras bases de datos para realizar los análisis respectivos.[1]

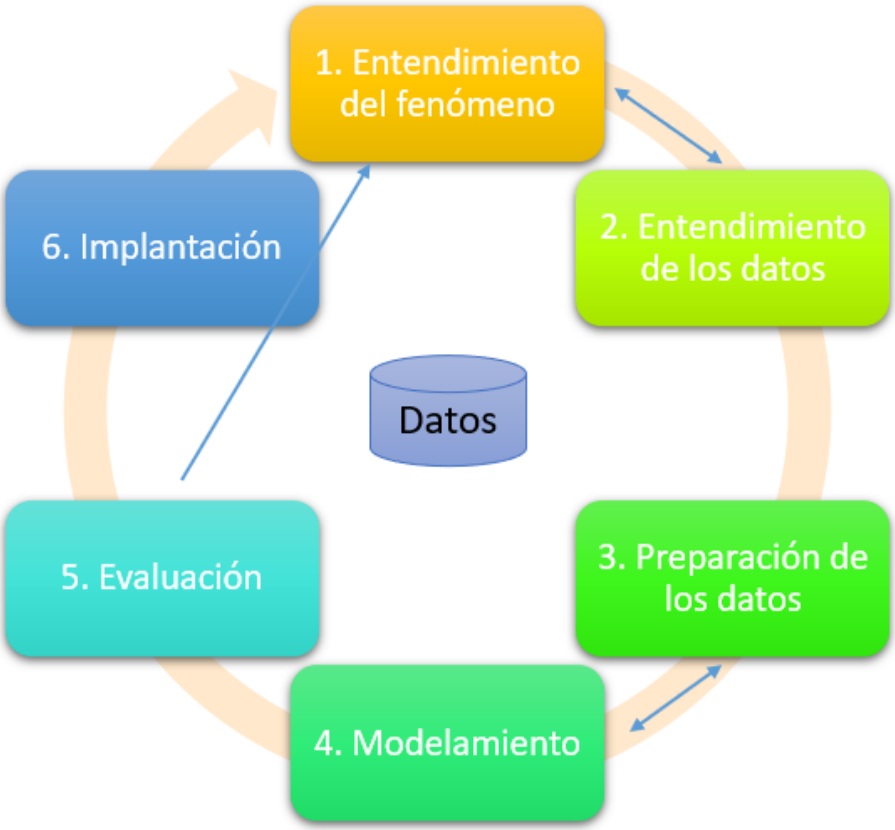


Figura 2. *Metodología CRISP-DM*. Fuente: Construcción propia.

A través de análisis factoriales exploratorios en el software SPSS, se realizarán cruces de variables según los hallazgos de la revisión de la literatura aquí expuesta que posibiliten la identificación de correlaciones, para posteriormente aplicar los resultados en la generación del modelo predictivo. Respecto a las variables analizadas, se tienen, las relacionadas con las bases de datos de las pruebas Saber 11 y las relacionadas con las pruebas Saber Pro.

Relacionadas con las bases de datos de las pruebas saber 11: género, edad, período de presentación del examen, pertenencia a una etnia, variables de discapacidad, departamento de

[1]Es importante agregar que los módulos de: lectura crítica, razonamiento cuantitativo, competencias ciudadanas, inglés y comunicación escrita, se componen de 177 preguntas de opción múltiple con única respuesta. De estas, 160 evalúan las competencias genéricas y las 17 restantes corresponden a un cuestionario socioeconómico con fines investigativos, por lo que no tiene calificación. En el módulo de comunicación escrita, hay una pregunta abierta donde se debe escribir un texto argumentativo. La duración de esta sesión es de 4 horas y 40 minutos, es decir que en promedio el estudiante cuenta con un minuto y medio por pregunta. La cantidad de preguntas por módulo son a saber:
•Razonamiento cuantitativo – 35
•Lectura crítica – 35
•Competencias ciudadanas – 35
•Inglés – 55
•Comunicación escrita – 1

residencia, estrato socioeconómico, ingreso familiar mensual, situación económica, género del colegio, naturaleza del colegio, carácter del establecimiento, puntaje en lectura crítica, puntaje en matemáticas, puntaje en ciencias naturales, puntaje en sociales y ciudadanas, puntaje inglés, puntaje total obtenido, índice socioeconómico.

Relacionadas con las bases de datos de las pruebas Saber Pro: género, edad, período de aplicación de la prueba, estado civil, pertenencia a una etnia, variables de discapacidad, departamento de residencia, estrato socioeconómico de la vivienda, ingreso familiar mensual, grupo de referencia al que pertenece el programa académico del estudiante, ingresos mensuales del hogar, grupo de referencia al que pertenece el programa académico del estudiante, departamento donde se ofrece el programa , metodología del programa académico, carácter académico de la IES, puntaje razonamiento cuantitativo, puntaje lectura crítica, puntaje competencias ciudadanas, puntaje de inglés, nivel de desempeño en inglés, puntaje comunicación escrita, desempeño comunicación escrita.

Las abreviaciones en la base de datos para las variables pueden consultarse en el Anexo 1.

Posterior a la identificación de correlaciones en los resultados de las pruebas mencionadas, se generarán las entradas del modelo, tomando como salida los resultados de la prueba Saber Pro para cada módulo, utilizando el software SPSS AMOS, en el cual se generan mediciones de variables latentes como una estrategia para medir índices de significancia estadística del modelo (Medidas de ajuste absoluto, medidas de ajuste incremental y medidas de ajuste de parsimonia)[2] (Escobedo et al., 2016).

Finalmente, se ensambla el modelo por nodos, y se generan nodos de origen para la base de datos integrada; la cual a su vez será filtrada por tipo y mediante un algoritmo clasificador automático de modelos, permitirá indicar la estructura conceptual que más se ajusta a las características de los datos relacionados entre las estructuras. Se tiene como opciones: clúster automático, árboles aleatorios, AS de árbol, lista de decisiones, C5.0, PCA/Factorial, red neuronal, red bayesiana, árbol XGBoost, CHAID y perceptrón multicapa.

A manera ilustrativa, el trabajo con algoritmos CHAID y redes neuronales permite la inclusión en una misma base de datos de variables cualitativas y cuantitativas. Además, al utilizar medidas de distancia entre los datos, estos posibilitan una contrastación de resultados que iterativamente mejoran las soluciones obtenidas.

En este sentido, la investigación mixta a utilizar es de estrategia concurrente de triangulación, según la cual: “un mismo estudio busca confirmar, correlacionar o corroborar. Utiliza alguna perspectiva teórica, en la interpretación busca la integración. Se recopilan datos cuantitativos y cualitativos simultáneamente” (Pereira, 2011).

[2] Bondad de ajuste absoluto: Determina el grado en que el modelo general predice la matriz de correlaciones y para SEM, el estadístico radio de verosimilitud Chi-cuadrado es la única medida estadística.
Medidas de ajuste incremental del modelo: estas medidas comparan el modelo propuesto con algún otro existente, llamado generalmente modelo nulo.
El Índice de bondad de ajuste de parsimonia (PGFI) es un índice que considera los grados de libertad disponibles para probar el modelo.

De la identificación y posterior aplicación de la opción más eficiente, se procede a identificar escenarios que interpretados en el contexto, permitirán apoyar la toma de decisiones en materia de gestión curricular en el ámbito nacional, departamental, municipal e institucional. Asimismo, los resultados se presentarán en forma Tableu a nivel dinámico, constituyéndose esta arquitectura en una base para la generación de un sistema de información que aporte a la toma de decisiones basadas en escenarios de modelos predictivos.

Para la consolidación de la base de datos, se ha desarrollado un diagrama integrado de datos que ha sido desarrollado en power pivot y que contiene los datos que se han de analizar.

En el anexo 2, se identifican las llaves que conectan los registros para las diferentes pruebas por período. Adicionalmente, utilizando el lenguaje de programación DAX, se realizan consultas a la base de datos para generar integraciones de datos que posibiliten la consolidación de la información para su posterior análisis en SPSS. Cada una de las ventanas que se presentan corresponden a hojas de cálculo de Excel que contienen todas las columnas de cada base de datos del ICFES para cada semestre de aplicación. Todas van a la hoja de cálculo denominada "Llaves". La llave es un identificador que se encuentra en la prueba Saber 11 y que se corresponde con un identificador en la prueba Saber Pro, de esa forma se puede hacer una consulta en para un registro específico en la base de datos. Por ejemplo, si un estudiante tiene consecutivo en las pruebas Saber 11 como "SABER1120062195960", entonces existe otro número en la prueba Saber Pro como "EK201210000532". Tal y como podrá observarse en el anexo 3 donde podrán identificarse las llaves de consecutivo de estudiante para las pruebas Saber 11 y las pruebas Saber Pro.

Con base en la información, se pueden integrar las bases de datos para hacer las consultas respectivas, tal mediante relaciones entre las llaves de las pruebas Saber 11 y las pruebas Saber Pro. La relación de la llave de pruebas Saber 11 se conjuga con la llave de las pruebas Saber Pro. De esa forma, es posible hacer una consulta de datos para posteriormente hacer el tratamiento estadístico que sea requerido. (ver anexo 3)

Resultados

Como aspecto clave para el desarrollo de la investigación, se consideró el desarrollo en primera instancia de la identificación de factores predictores de las pruebas Saber Pro, tanto en el puntaje global como en los módulos de razonamiento cuantitativo, lectura crítica, competencia ciudadana, inglés y comunicación escrita. Para tal efecto, se consideraron todas las variables, tanto las cualitativas como las cuantitativas y se utilizó un algoritmo de selección de características el cual permite realizar un cribado de predictores, en ese sentido,

> "ayuda a identificar los campos que son más importantes para predecir determinados resultados. De un conjunto de cientos e incluso miles de predictores, el nodo Selección de características, filtra, ordena por rango y selecciona los predictores que pueden ser más importantes. En última instancia, puede lograr un modelo más eficaz y rápido, que utilice menos predictores, se ejecute de manera más rápida y sea más fácil de entender". (IBM, 2021, pág. 84)

Con base en el proceso de cribado, se seleccionan los registros y se genera una ruta de analítica, que considere todas las entradas posibles que se quiera tener en cuenta a manera de entrada del

modelo y teniendo en cuenta una salida. Un primer análisis que se hizo con la finalidad de identificar la fiabilidad del modelo, fue incluir también las variables determinantes del nivel de desempeño de saber Pro, como variables de entrada, tal y como se podrá observar en el anexo 4.

El desarrollo de este algoritmo se conjuga con el modelo CHAID generando la red de analítica de la Figura 3.

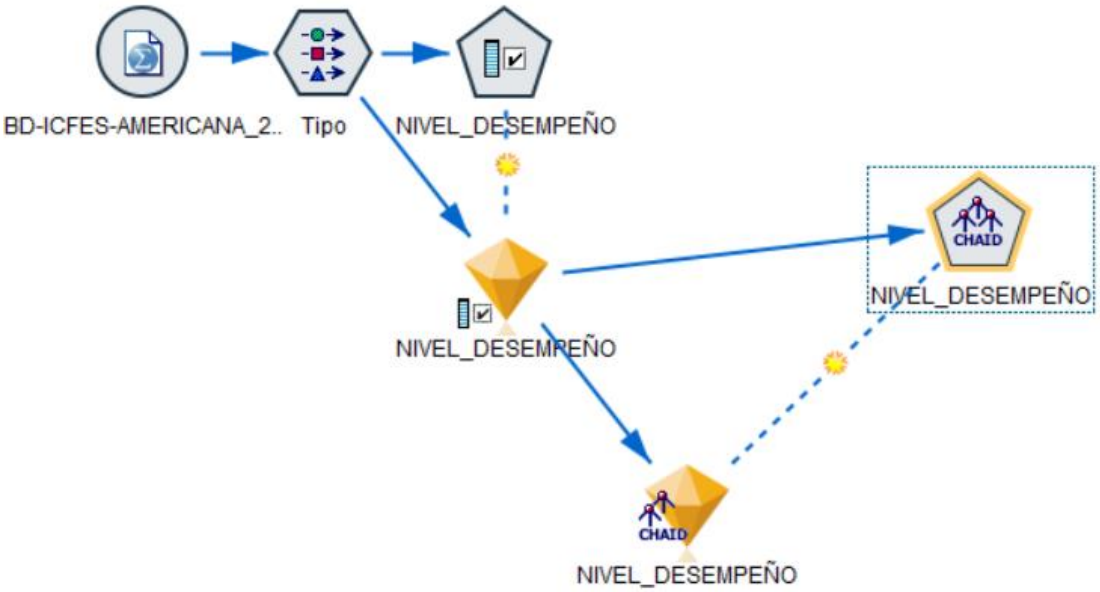


Figura 3. Relaciones para el cribado de variables, etapa exploratoria.

El algoritmo CHAID “es un método de clasificación que genera árboles de decisión utilizando un tipo específico de estadísticos denominados estadísticos chi-cuadrado para determinar los mejores lugares para realizar las divisiones en el árbol de decisión.” (IBM, 2021). Con base en los campos, se identificaron las variables predictoras de las Figura 4.

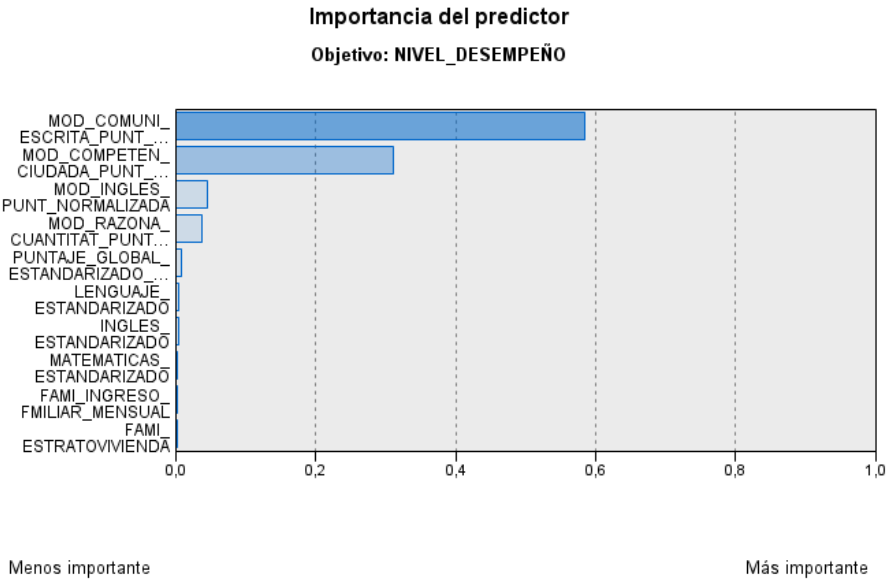


Figura 4. Importancia del predictor en el análisis exploratorio.

Naturalmente, el desarrollo de esta etapa exploratoria produjo en mayores niveles de importancia predictiva a las variables de los módulos de las pruebas Saber Pro, los cuales son los determinantes del puntaje global. Se resalta que bajo esta configuración, la comunicación escrita y la competencia ciudadana cuentan con los valores más altos de incidencia en el cálculo del resultado. Teniendo en cuenta lo anterior, se procedió a retirar las variables asociadas a los módulos y se incluyeron variables sociodemográficas, así como variables relacionadas con los puntajes de las pruebas saber 11, obteniendo los resultados que se presentan en la Figura 5.

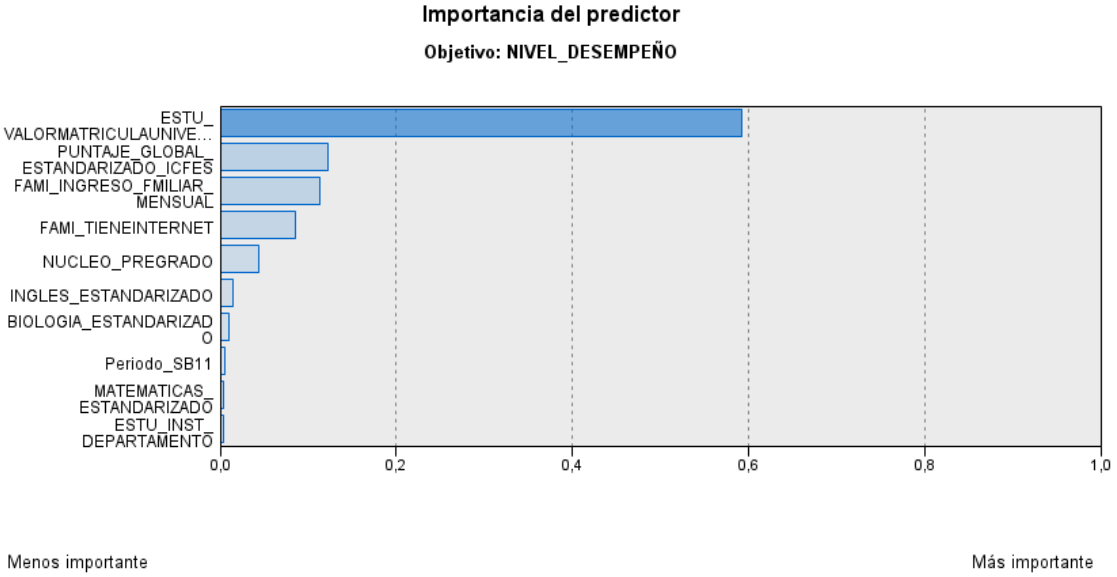


Figura 5. Importancia de predictores de los niveles de desempeño en la prueba saber Pro para los años 2012 a 2019.

En la Figura 5 se identifican según nivel de importancia predictiva el rango de valor de matrícula de la universidad, el puntaje global Saber 11, el ingreso familiar mensual, si la familia tiene o no internet, el núcleo de pregrado, el puntaje de inglés en la prueba Saber 11, el puntaje de biología o ciencias naturales en la prueba Saber 11, el período en el cual se presentó la prueba Saber 11, el puntaje de la prueba de matemáticas Saber 11 y por último el departamento del cual es la Institución de Educación Superior.

Lo anterior se explica al identificar que para la calificación de las pruebas, se emplea la Teoría de Respuesta al Ítem (TRI), que "permite estimar la habilidad de los individuos (θ) con base en las respuestas dadas a los ítems y las características de los ítems (a los cuales llamamos parámetros)" (ICFES, 2018). Actualmente, se utiliza un modelo logístico de tres parámetros (3PL por sus siglas en inglés) en el cual se tiene en cuenta, la dificultad del ítem (b), la discriminación o posibilidad de diferenciación entre resultados (a) y la probabilidad de acierto casual (c), bajo el supuesto que "al analizar la relación entre la habilidad y la probabilidad de acierto, cabe señalar que estas no tienen un comportamiento lineal, lo que quiere decir que un aumento en una unidad en habilidad no implica un aumento proporcional en la probabilidad de acertar el ítem. En tal caso, se recurre a una función matemática que permita expresar la no linealidad de la relación y que además tenga en cuenta que la probabilidad de acierto del ítem está entre 0 y 1". (Pág. 3)

La ecuación 1, presenta el cálculo de los puntajes en el modelo 3PL para el cálculo de las puntuaciones en las pruebas ICFES, en este sentido se puede notar que se utiliza una función de activación, similar a la que se utilizaría en el tratamiento de redes neuronales en analítica de datos.

$$P(X_j = 1 \mid \theta,a,b,c,) = P(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1+e^{1.7a_i(\theta - b_i)}} \quad (1)$$

Con lo anterior se identifica que las pruebas Saber 11 son buenos predictores de las pruebas Saber Pro, sin embargo, se identifican variables socioeconómicas que también pueden estar implicadas como el rango de valor del costo de matrícula.

Al identificar las variables, se aplicaron en diferentes algoritmos que permitieron la generación de modelos. Un primer modelo utilizó como entradas los resultados de las pruebas Saber 11, tal y como lo propone el ICFES, utilizando un modelo de regresión logística multivariada, se obtuvo los resultados de la Tabla 2.

Tabla 2. Estimaciones de parámetro para el modelo de regresión logística.

							95% de intervalo de confianza para Exp(B)	
NIVEL_DESEMPEÑO	B	Desv. Error	Wald	Gl	Sig.	Exp(B)	Límite inferior	Límite superior

1	Intersección	5,513	0,052	11082,655	1	0,000			
	LENGUAJE_ESTANDARIZADO	-1,31	0,088	221,118	1	<0,001	0,270	0,227	0,320
	INGLES_ESTANDARIZADO	-2,96	0,080	1365,537	1	<0,001	0,052	0,044	0,060
	MATEMATICAS_ESTANDARIZADO	-2,32	0,091	650,548	1	<0,001	0,098	0,082	0,117
	BIOLOGIA_ESTANDARIZADO	-1,12	0,085	175,215	1	<0,001	0,325	0,275	0,384
	CIENCIAS_SOCIALES_ESTANDARIZADO	-1,72	0,095	331,238	1	<0,001	0,178	0,147	0,214
2	Intersección	5,248	0,047	12551,796	1	0,000			
	LENGUAJE_ESTANDARIZADO	-,560	0,080	48,920	1	<0,001	0,571	0,488	0,668
	INGLES_ESTANDARIZADO	-2,85	0,069	1700,928	1	0,000	0,058	0,050	0,066
	MATEMATICAS_ESTANDARIZADO	-1,39	0,082	287,872	1	<0,001	0,247	0,210	0,291
	BIOLOGIA_ESTANDARIZADO	-,425	0,076	30,829	1	<0,001	0,654	0,563	0,760
	CIENCIAS_SOCIALES_ESTANDARIZADO	-,446	0,086	26,716	1	<0,001	0,640	0,541	0,758
3	Intersección	3,747	0,045	6879,626	1	0,000			
	LENGUAJE_ESTANDARIZADO	0,252	0,078	10,488	1	0,001	1,287	1,105	1,500
	INGLES_ESTANDARIZADO	-1,13	0,067	290,193	1	<0,001	0,322	0,282	0,367
	MATEMATICAS_ESTANDARIZADO	-,164	0,080	4,214	1	0,040	0,848	0,725	0,993
	BIOLOGIA_ESTANDARIZADO	0,075	0,074	1,017	1	0,313	1,078	0,932	1,247
	CIENCIAS_SOCIALES_ESTANDARIZADO	0,478	0,084	32,545	1	<0,001	1,613	1,369	1,902
4	Intersección	1,298	0,046	788,070	1	<0,001			
	LENGUAJE_ESTANDARIZADO	0,568	0,080	50,336	1	<0,001	1,765	1,508	2,064
	INGLES_ESTANDARIZADO	0,220	0,068	10,516	1	0,001	1,246	1,091	1,424
	MATEMATICAS_ESTANDARIZADO	0,386	0,082	22,122	1	<0,001	1,471	1,252	1,727
	BIOLOGIA_ESTANDARIZADO	0,037	0,076	,237	1	0,627	1,038	0,894	1,204
	CIENCIAS_SOCIALES_ESTANDARIZADO	1,112	0,086	166,476	1	<0,001	3,042	2,569	3,602

En la Tabla, las comas representan separador decimal. Respecto a la abreviatura gl, se refiere a grado de libertad, sig. Implica significancia estadística.

Se identifica entonces una relación funcional entre los resultados de las pruebas saber 11 y las pruebas saber Pro, con niveles de significancia altos en la mayoría de los componentes.

Con base en lo anterior, y con la finalidad de considerar posibles relaciones con variables socioeconómicas, se generó un modelo de redes neuronales[3], tal y como se presenta en la Figura 6.

Para la clasificación de puntuaciones en la prueba Saber Pro, se utilizaron los datos de la Tabla 3:

Tabla 3. Puntuaciones por nivel en la prueba Saber Pro.

Puntuación	Nivel
0 – 124	1

[3] El modelo de red neuronal se realizó con semilla 123456789 con la finalidad de obtener replicabilidad en los resultados.

125 – 149	2
150 – 199	3
200 – 300	4

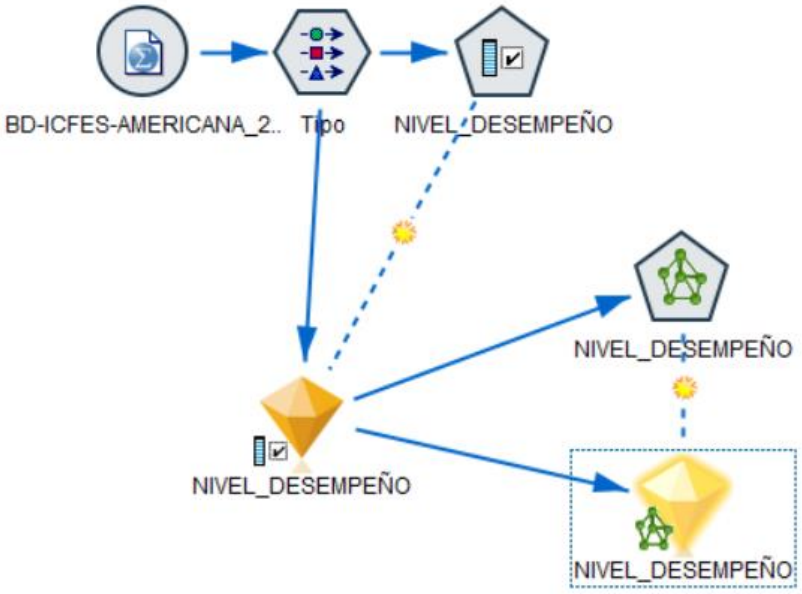


Figura 6. Modelo de redes neuronales que integra variables cualitativas y cuantitativas.

El modelo generó el resultado de la Figura 7, verificando la importancia predictora de las variables anteriormente identificadas.

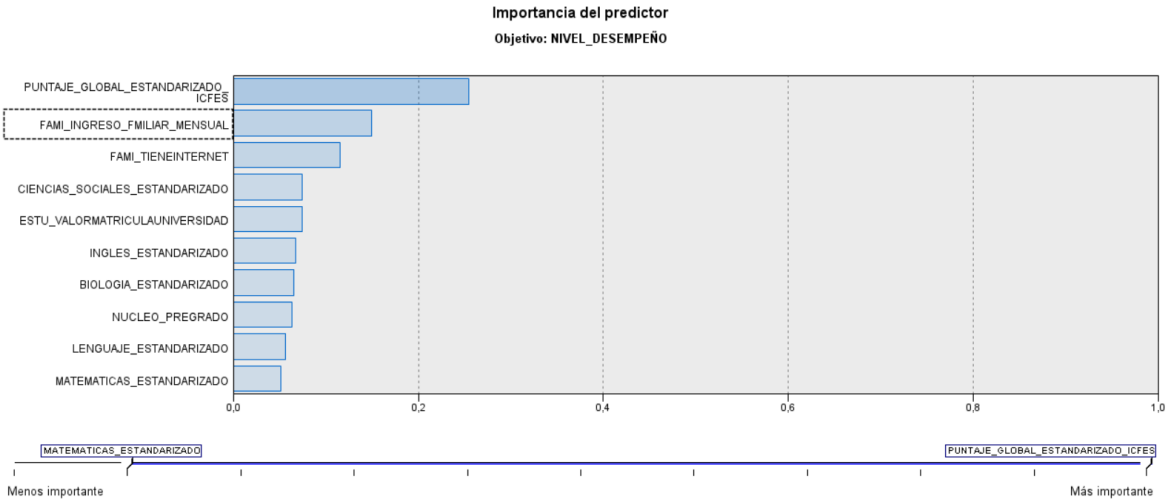


Figura 7. Estimación de la Prueba Saber Pro explicado por Saber 11 y variables sociodemográficas bajo el modelo de redes neuronales

Se identifica entonces la incidencia de variables sociodemográficas en los resultados de las pruebas Saber Pro, como una forma de reinterpretar las dinámicas sociales en el marco de los procesos académicos en las Instituciones de Educación Superior. Se procedió por lo tanto a realizar nuevamente un análisis logístico multivariado incluyendo las variables identificadas en la Figura 7 que son, el valor de matrícula universitaria y el hecho de tener o no internet en la vivienda, con lo cual se obtienen los resultados de la Tabla 4.

Tabla 4. Estimaciones de parámetro para variables sociodemográficas y puntaciones en las pruebas Saber 11.

NIVEL_DESEMPEÑO[a]		B	Desv. Error	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B)	
								Límite inferior	Límite superior
0	Intersección	1,607	0,149	115,923	1	<0,001			
	LENGUAJE_ESTANDARIZADO	2,199	1,383	2,528	1	0,112	9,016	0,600	135,588
	INGLES_ESTANDARIZADO	-2,81	0,493	32,527	1	<0,001	0,060	0,023	0,158
	MATEMATICAS_ESTANDARIZADO	0,825	1,372	0,362	1	0,547	2,282	0,155	33,566
	BIOLOGIA_ESTANDARIZADO	2,526	1,385	3,324	1	0,068	12,503	0,827	188,918
	CIENCIAS_SOCIALES_ESTANDARIZADO	2,482	1,390	3,189	1	0,074	11,966	0,785	182,408
	PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES	-13,9	5,921	5,560	1	0,018	8,642E-7	7,8E-12	0,095
	[ESTU_VALORMATRICULAUNIVERSIDAD=]	-1,95	0,105	347,417	1	<0,001	0,142	0,115	0,174
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones]	1,543	0,094	266,982	1	<0,001	4,679	3,889	5,631
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones]	0,412	0,085	23,673	1	<0,001	1,510	1,279	1,782
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones]	1,582	0,100	251,317	1	<0,001	4,863	4,000	5,914
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones]	0,458	0,106	18,829	1	<0,001	1,581	1,286	1,945
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones]	1,794	0,120	223,515	1	<0,001	6,013	4,753	7,608
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones]	2,080	0,152	187,276	1	<0,001	8,001	5,940	10,777
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón]	0,626	0,089	49,039	1	<0,001	1,869	1,569	2,227
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones]	0,004	0,119	0,001	1	0,0976	1,004	0,794	1,268

	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones]	2,389	0,192	154,497	1	<0,0001	10,907	7,483	15,897
	[ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones]	2,094	0,235	79,126	1	<0,0001	8,121	5,119	12,884
	[ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil]	0,539	0,090	36,176	1	<0,0001	1,714	1,438	2,043
	[ESTU_VALORMATRICULAUNIVERSIDAD=No pagó matrícula]	0	0	.	0	.	.	.	.
	[FAMI_TIENEINTERNET=]	2,217	0,060	1372,537	1	<0,001	9,179	8,163	10,321
	[FAMI_TIENEINTERNET=No]	,264	0,073	13,121	1	<0,001	1,302	1,129	1,501
	[FAMI_TIENEINTERNET=Si]	0	.	.	0	.	.	.	.
1	Intersección	6,335	0,079	6493,836	1	0,000			
	LENGUAJE_ESTANDARIZADO	-,390	0,475	0,673	1	0,412	0,677	0,267	1,718
	INGLES_ESTANDARIZADO	-3,61	0,191	359,144	1	<0,001	0,027	0,019	0,039
	MATEMATICAS_ESTANDARIZADO	-1,44	0,458	9,887	1	0,002	0,237	0,097	0,582
	BIOLOGIA_ESTANDARIZADO	0,108	0,477	0,051	1	0,821	1,114	0,437	2,839
	CIENCIAS_SOCIALES_ESTANDARIZADO	-0,82	0,479	2,917	1	0,088	0,442	0,173	1,128
	PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES	-9,01	1,994	20,440	1	<0,001	0,000	2,448E-6	0,006
	[ESTU_VALORMATRICULAUNIVERSIDAD=]	-2,14	0,073	867,062	1	<0,001	0,118	0,102	0,136
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones]	3,312	0,055	3599,444	1	0,000	27,440	24,626	30,576
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones]	0,054	0,067	0,646	1	0,422	1,055	0,926	1,202
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones]	3,050	0,057	2825,119	1	0,000	21,111	18,865	23,623
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones]	0,515	0,086	35,806	1	<0,001	1,674	1,414	1,982
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones]	2,886	0,068	1813,527	1	0,000	17,926	15,696	20,473
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones]	3,068	0,086	1277,150	1	<0,001	21,497	18,168	25,436
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón]	2,045	0,056	1356,571	1	<,0001	7,731	6,933	8,619
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones]	0,677	0,090	56,420	1	<,0001	1,968	1,649	2,348
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones]	3,680	0,104	1252,971	1	<,0001	39,643	32,335	48,602

	[ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones]	2,743	0,124	490,983	1	<0,001	15,541	12,192	19,809
	[ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil]	2,126	0,055	1468,992	1	0,000	8,380	7,517	9,342
	[ESTU_VALORMATRICULAUNIVERSIDAD=No pagó matrícula]	0	.	.	0	.	.	.	.
	[FAMI_TIENEINTERNET=]	-0,28	0,040	48,334	1	<0,001	0,755	0,697	0,817
	[FAMI_TIENEINTERNET=No]	0,419	0,031	184,750	1	<0,001	1,520	1,431	1,614
	[FAMI_TIENEINTERNET=Si]	0	.	.	0	.	.	.	.
2	Intersección	6,412	0,066	9372,739	1	0,000			
	LENGUAJE_ESTANDARIZADO	1,840	0,335	30,089	1	<0,001	6,298	3,263	12,155
	INGLES_ESTANDARIZADO	-3,20	0,144	497,959	1	<0,001	0,041	0,031	0,054
	MATEMATICAS_ESTANDARIZADO	,605	0,315	3,691	1	0,055	1,831	0,988	3,393
	BIOLOGIA_ESTANDARIZADO	2,254	0,338	44,575	1	<0,001	9,521	4,913	18,450
	CIENCIAS_SOCIALES_ESTANDARIZADO	1,708	0,339	25,385	1	<0,001	5,519	2,840	10,727
	PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES	-13,6	1,381	98,047	1	<0,001	1,154E-6	7,710E-8	1,728E-5
	[ESTU_VALORMATRICULAUNIVERSIDAD=]	-2,47	0,039	4027,444	1	0,000	0,084	0,078	0,091
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones]	1,919	0,044	1873,499	1	0,000	6,811	6,244	7,429
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones]	0,046	0,046	1,026	1	0,311	1,047	0,958	1,145
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones]	1,995	0,046	1877,524	1	0,000	7,354	6,720	8,049
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones]	-,216	0,060	12,835	1	<0,001	0,806	0,716	0,907
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones]	2,132	0,055	1484,744	1	0,000	8,429	7,563	9,394
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones]	2,368	0,072	1089,124	1	<0,001	10,672	9,272	12,283
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón]	0,653	0,044	218,967	1	<0,001	1,921	1,762	2,094
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones]	-,834	0,061	188,880	1	<0,001	0,434	0,386	0,489
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones]	2,755	0,088	983,986	1	<0,001	15,728	13,241	18,683
	[ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones]	1,836	0,096	366,748	1	<0,001	6,273	5,198	7,569
	[ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil]	0,565	0,044	165,864	1	<0,001	1,759	1,614	1,917

	[ESTU_VALORMATRICULAUNIVERSIDAD=No pagó matrícula]	0	0.	.	0	.	.	.	.
	[FAMI_TIENEINTERNET=]	2,016	0,028	5273,141	1	0,000	7,508	7,111	7,928
	[FAMI_TIENEINTERNET=No]	0,113	0,029	15,607	1	<0,001	1,119	1,059	1,184
	[FAMI_TIENEINTERNET=Si]	0	.	.	0	.	.	.	.
3	Intersección	4,159	0,063	4354,765	1	0,000			
	LENGUAJE_ESTANDARIZADO	1,065	0,240	19,661	1	<0,001	2,902	1,812	4,648
	INGLES_ESTANDARIZADO	-1,01	0,118	73,684	1	<0,001	0,363	0,288	0,457
	MATEMATICAS_ESTANDARIZADO	0,842	0,211	15,873	1	<0,001	2,320	1,534	3,511
	BIOLOGIA_ESTANDARIZADO	1,040	0,243	18,363	1	<0,001	2,831	1,759	4,556
	CIENCIAS_SOCIALES_ESTANDARIZADO	1,702	0,245	48,465	1	<0,001	5,487	3,398	8,861
	PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES	-4,12	0,945	19,100	1	<0,001	0,016	0,003	0,103
	[ESTU_VALORMATRICULAUNIVERSIDAD=]	-2,19	0,035	3975,834	1	0,000	0,111	0,104	0,119
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones]	-1,29	0,043	902,239	1	<0,001	0,273	0,251	0,297
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones]	0,411	0,044	86,485	1	<0,001	1,509	1,383	1,645
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones]	-,984	0,045	483,360	1	<0,001	0,374	0,342	0,408
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones]	0,518	0,058	79,753	1	<0,001	1,678	1,498	1,881
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones]	-0,51	0,054	92,632	1	<0,001	0,597	0,537	0,663
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones]	-0,13	0,070	3,979	1	0,046	0,870	0,759	0,998
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón]	-0,36	0,043	73,272	1	<0,001	0,694	0,638	0,755
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones]	0,386	0,056	46,837	1	<0,001	1,472	1,318	1,644
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones]	0,828	0,085	95,568	1	<0,001	2,289	1,939	2,703
	[ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones]	-0,16	0,089	3,269	1	0,071	0,852	0,716	1,014
	[ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil]	-0,24	0,042	33,176	1	<,0001	0,784	0,721	0,852
	[ESTU_VALORMATRICULAUNIVERSIDAD=No pagó matrícula]	0	0.	.	0	0.	.	.	.
	[FAMI_TIENEINTERNET=]	-,069	0,027	6,836	1	0,009	0,933	0,886	0,983
	[FAMI_TIENEINTERNET=No]	0,242	0,027	80,040	1	<0,001	1,273	1,208	1,343

	[FAMI_TIENEINTERNET=Si]	0	.	.	0	.	.	.	.
4	Intersección	-,651	0,067	95,164	1	<0,001			
	LENGUAJE_ESTANDARIZADO	-,593	0,248	5,721	1	0,017	0,553	0,340	0,898
	INGLES_ESTANDARIZADO	,665	0,123	29,400	1	<0,001	1,944	1,529	2,473
	MATEMATICAS_ESTANDARIZADO	,086	0,216	0,160	1	0,690	1,090	0,714	1,664
	BIOLOGIA_ESTANDARIZADO	-1,02	0,250	16,576	1	<0,001	0,361	0,221	0,589
	CIENCIAS_SOCIALES_ESTANDARIZADO	1,275	0,253	25,459	1	<0,001	3,580	2,181	5,874
	PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES	7,720	0,967	63,691	1	<0,001	2253,006	338,346	15002,501
	[ESTU_VALORMATRICULAUNIVERSIDAD=]	-2,77	0,039	5117,810	1	0,000	0,062	0,058	0,067
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones]	-5,10	0,062	6804,626	1	0,000	0,006	0,005	0,007
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones]	0,702	0,045	244,099	1	<0,001	2,019	1,848	2,205
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones]	-5,30	0,070	5788,497	1	0,000	0,005	0,004	0,006
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones]	1,015	0,059	298,271	1	<0,001	2,759	2,459	3,096
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones]	-4,78	0,074	4192,411	1	0,000	0,008	0,007	0,010
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones]	-4,05	0,084	2328,541	1	0,000	0,017	0,015	0,020
	[ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón]	-0,80	0,044	334,481	1	<0,001	0,449	0,412	0,489
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones]	1,238	0,057	466,841	1	<0,001	3,449	3,083	3,859
	[ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones]	-2,47	0,088	789,160	1	<0,001	0,084	0,071	0,100
	[ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones]	-3,28	0,095	1195,520	1	<0,001	0,037	0,031	0,045
	[ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil]	-,612	0,043	200,383	1	<0,001	0,542	0,498	0,590
	[ESTU_VALORMATRICULAUNIVERSIDAD=No pagó matrícula]	0	.	.	0	0.	.	.	.
	[FAMI_TIENEINTERNET=]	-4,42	0,038	13843,407	1	0,000	0,012	0,011	0,013
	[FAMI_TIENEINTERNET=No]	-0,03	0,028	1,023	1	0,312	0,972	0,921	1,027
	[FAMI_TIENEINTERNET=Si]	0	.	.	0	.	.	.	.

En la Tabla, las comas representan separador decimal. Respecto a la abreviatura gl, se refiere a grado de libertad, sig. Implica significancia estadística.

Teniendo en cuenta que los resultados de la prueba Saber Pro, se categorizaron en los grupos de la Tabla anterior, se obtienen las siguientes ecuaciones de interpretación de los modelos:

Ecuación para 1

-0,39 * LENGUAJE_ESTANDARIZADO +
-3,616 * INGLES_ESTANDARIZADO +
-1,439 * MATEMATICAS_ESTANDARIZADO +
0,1083 * BIOLOGIA_ESTANDARIZADO +
-0,8176 * CIENCIAS_SOCIALES_ESTANDARIZADO +
-9,013 * PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES +
-2,141 * [ESTU_VALORMATRICULAUNIVERSIDAD=] +
3,312 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones] +
0,05357 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones] +
3,05 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones] +
0,5152 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones] +
2,886 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones] +
3,068 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones] +
2,045 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón] +
0,6768 * [ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones] +
3,68 * [ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones] +
2,743 * [ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones] +
2,126 * [ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil] +
-0,2811 * [FAMI_TIENEINTERNET=] +
0,4185 * [FAMI_TIENEINTERNET=No] +
+ 6,335

Ecuación para 2

1,84 * LENGUAJE_ESTANDARIZADO +
-3,205 * INGLES_ESTANDARIZADO +
0,6047 * MATEMATICAS_ESTANDARIZADO +
2,254 * BIOLOGIA_ESTANDARIZADO +
1,708 * CIENCIAS_SOCIALES_ESTANDARIZADO +
-13,67 * PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES +
-2,479 * [ESTU_VALORMATRICULAUNIVERSIDAD=] +
1,919 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones] +
0,04626 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones] +
1,995 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones] +
-0,216 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones] +
2,132 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones] +
2,368 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones] +
0,6528 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón] +
-0,8341 * [ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones] +
2,755 * [ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones] +
1,836 * [ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones] +
0,5647 * [ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil] +
2,016 * [FAMI_TIENEINTERNET=] +
0,1128 * [FAMI_TIENEINTERNET=No] +
+ 6,412

Ecuación para 3

1,065 * LENGUAJE_ESTANDARIZADO +
-1,015 * INGLES_ESTANDARIZADO +
0,8417 * MATEMATICAS_ESTANDARIZADO +
1,04 * BIOLOGIA_ESTANDARIZADO +
1,702 * CIENCIAS_SOCIALES_ESTANDARIZADO +
-4,129 * PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES +
-2,199 * [ESTU_VALORMATRICULAUNIVERSIDAD=] +
-1,299 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones] +
0,4112 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones] +
-0,9843 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones] +
0,5179 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones] +
-0,5166 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones] +
-0,1389 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones] +
-0,3652 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón] +
0,3865 * [ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones] +
0,8283 * [ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones] +
-0,1603 * [ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones] +
-0,2436 * [ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil] +
-0,06941 * [FAMI_TIENEINTERNET=] +
0,2416 * [FAMI_TIENEINTERNET=No] +
+ 4,159

Ecuación para 4

-0,5928 * LENGUAJE_ESTANDARIZADO +
0,6649 * INGLES_ESTANDARIZADO +
0,0862 * MATEMATICAS_ESTANDARIZADO +
-1,02 * BIOLOGIA_ESTANDARIZADO +
1,275 * CIENCIAS_SOCIALES_ESTANDARIZADO +
7,72 * PUNTAJE_GLOBAL_ESTANDARIZADO_ICFES +
-2,779 * [ESTU_VALORMATRICULAUNIVERSIDAD=] +
-5,1 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 2.5 millones] +
0,7025 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 1 millón y menos de 3 millones] +
-5,3 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 2.5 millones y menos de 4 millones] +
1,015 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 3 millones y menos de 5 millones] +
-4,787 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 4 millones y menos de 5.5 millones] +
-4,058 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 5.5 millones y menos de 7 millones] +
-0,8001 * [ESTU_VALORMATRICULAUNIVERSIDAD=Entre 500 mil y menos de 1 millón] +
1,238 * [ESTU_VALORMATRICULAUNIVERSIDAD=Más de 5 millones] +
-2,474 * [ESTU_VALORMATRICULAUNIVERSIDAD=Más de 7 millones] +
-3,284 * [ESTU_VALORMATRICULAUNIVERSIDAD=Mas de 7 millones] +
-0,6123 * [ESTU_VALORMATRICULAUNIVERSIDAD=Menos de 500 mil] +
-4,425 * [FAMI_TIENEINTERNET=] +
-0,02819 * [FAMI_TIENEINTERNET=No] +
+ -0,6515

Discusión

La presente investigación tuvo dentro de sus objetivos la posibilidad de proporcionar a directores de programas y docentes, mediante un análisis de información, la capacidad de apoyar de manera propositiva los programas educativos en sus instituciones, logrando predecir el rendimiento de los estudiantes, no solo a través de sus notas sino utilizando factores predictivos como datos sociodemográficos.

Respecto a los resultados, ha sido posible obtener a manera de conclusión la identificación de determinantes que explican mayormente los resultados de las pruebas Saber Pro, éstos son los puntajes en los módulos de comunicación escrita y de competencia ciudadana. Respecto a los modelos predictivos, se ha identificado que efectivamente las pruebas Saber 11 permiten identificar tendencias de predicción de los resultados de las pruebas Saber Pro tal y como se menciona en ICFES (2018). Un aspecto relevante del estudio realizado es con la aplicación de los algoritmos de predicción, fue posible identificar la existencia de importancia predictiva de variables sociodemográficas asociadas al ingreso familiar mensual y la conectividad a internet, de lo anterior se ratifica la importancia del cálculo actual del Índice de Nivel Socioeconómico - INSE (ICFES, 2019), como un factor predictor que podría articularse a estudios posteriores.

Los modelos predictivos en la educación son de creciente interés para los actores de las organizaciones educativas tanto en Colombia como en todo el mundo. Las instituciones educativas buscan encontrar nuevas oportunidades de análisis en los rastros digitales y en las grandes cantidades de información que acumulan los estudiantes en su proceso de enseñanza y aprendizaje. Las ecuaciones anteriores pueden ser utilizadas para interpretar las diferentes incidencias de los resultados de las pruebas saber 11 y las variables socioeconómicas según los

niveles de las pruebas. El objetivo es proporcionar a los estudiantes vías de aprendizaje más individualizadas y entornos de aprendizaje más adaptables que tengan en cuenta las necesidades de cada alumno. A su vez, se alienta a los maestros a utilizar una variedad de herramientas para apoyar el aprendizaje y brindar orientación oportuna, entre ellas la minería de datos.

En este sentido, el análisis de datos educativos (Educational Data Mining) (ADM) es la línea de investigación relacionada con la aplicación de métodos de minería de datos, aprendizaje automático y estadísticas a la información producida por Instituciones,Hernández-Leal, E. J., et. al (2018). A nivel del funcionamiento efectivo del sistema de gestión de la calidad de una institución, es necesario crear información entorno a que luego se permita gestionar el proceso de recopilación y análisis de datos, es por ello necesario incorporar métodos de procesamiento de datos en las tareas de gestión de la calidad del proceso educativo, uno de estos métodos es la minería de datos Soltan, G. Zh., et.al, (2013)

El uso generalizado del e-learning sistémico (e-learning) permite a los investigadores una cantidad mucho mayor de información en comparación con los procesos educativos. Esto se debe tanto al uso activo de herramientas digitales y varias tecnologías de recopilación de datos y con una audiencia más amplia en entornos educativos electrónicos. El crecimiento de los datos ha impulsado la aparición de una nueva dirección en el campo de la inteligencia artificial, la cual es el análisis de datos educativos, como lo menciona Baker, R. S., et. al, (2014). Para la gestión de las organizaciones educativas, los modelos predictivos parecen ser una oportunidad para desarrollar la educación, mejorar el progreso del estudio y prevenir la deserción.

Debido a la digitalización de los procesos académicos, las universidades están generando una gran cantidad de datos pertenecientes a estudiantes en formato electrónico. Es crucial para ellos transformar de manera efectiva esta recopilación masiva de datos en conocimiento que ayudará a los maestros, administradores y los responsables de la formulación de políticas para analizarlo para mejorar la toma de decisiones (Rodrigues y Zarate, 2018).

Los desarrollos en este campo se mueven cada vez más en la dirección de la analítica predictiva. Al analizar los datos acumulados y utilizar algoritmos basados en ciertos indicadores, es posible predecir, por ejemplo, el éxito y el progreso del estudio de los estudiantes, así como la finalización y la posible interrupción en el tiempo objetivo. Pero, ¿cuáles son los desafíos éticos involucrados en el uso de modelos predictivos? ¿Se pueden sellar las predicciones por adelantado por un estudiante o grupo de estudiantes en particular? ¿Qué efectos pueden tener las predicciones sobre la motivación o las acciones de estudiantes y profesores? ¿Es posible tener en cuenta los factores que afectan al alumno individual y al aprendizaje de manera suficientemente amplia en el pronóstico? ¿Qué tan confiables pueden ser los pronósticos? ¿Pueden las predicciones volverse autocumplidas?, la aplicación de modelos predictivos en la educación es en la actualidad todo un reto.

Por otro lado, el desarrollo de la analítica del aprendizaje en los últimos años ha sido significativo y se han lanzado programas piloto en varias organizaciones educativas diferentes. Sin embargo, cuando se trata de trabajo de desarrollo, es bueno recordar que las medidas relacionadas con la analítica del aprendizaje siempre deben implementarse teniendo en cuenta consideraciones éticas. Es esencial pensar detenidamente de antemano qué tipo de conocimiento y análisis tienen un papel real en el apoyo al aprendizaje y cómo se utilizarán para apoyar y promover el aprendizaje.

Es significativo considerar que los algoritmos predictivos deben usarse como parte de la analítica de aprendizaje con mucho cuidado y teniendo en cuenta ciertas condiciones límite. Debe recordarse que los modelos predictivos se basan solo en los datos disponibles (por ejemplo, datos de antecedentes, rutas de aprendizaje). Estos datos suelen ser una medida e indicador muy limitados de lo que es realmente un estudiante, cómo trabaja y qué cosas afectan su vida en un momento dado. Por lo tanto, las predicciones no pueden, en principio, proporcionar una imagen general del estudiante o de todos los factores y motivos individuales asociados con el estudiante.

Los modelos que predicen el rendimiento de los estudiantes siempre se basan en los únicos datos disponibles actualmente y las predicciones no se "mejoran" en ninguna etapa. También debe tenerse en cuenta que las predicciones, especialmente desde una perspectiva individual, a menudo son bastante inciertas y no permiten sacar conclusiones de gran alcance sobre el alumno individual. (Van Staalduinen et al.2018). Sin embargo, además de varios desafíos, el análisis predictivo también tiene muchas oportunidades, especialmente desde la perspectiva del desarrollo de la educación, la enseñanza y la orientación.

El potencial de la analítica predictiva se puede aprovechar mejor si se tienen en cuenta los desafíos éticos y las perspectivas mencionadas anteriormente al planificar el uso de la analítica del aprendizaje. Es particularmente importante considerar a qué se dirige el uso de modelos predictivos y para qué se utilizan. ¿Cómo apoyarán las proyecciones a los estudiantes y sus necesidades en las diferentes etapas de sus estudios? ¿Qué medidas se tomarán después de las previsiones? Por ejemplo, ¿los pronósticos permiten ver cosas que son importantes y relevantes para el éxito en los estudios? ¿Sobre qué base se puede desarrollar la educación y promover el aprendizaje? ¿Brindan la oportunidad de actuar y realizar las intervenciones de orientación necesarias en función de los comentarios recibidos?

Se ha descubierto que el análisis predictivo es particularmente útil cuando proporciona información interpretable para los pronósticos. Los intérpretes deben conocer adecuadamente el contexto de los datos recopilados y otros factores que afectan la interpretación. También deben tener suficiente experiencia en la selección, el momento y la implementación de las intervenciones de orientación necesarias. El informe afirma que, entre otras cosas, las visualizaciones versátiles permiten apoyar la interpretación de modelos predictivos al resaltar las variables que han influido en las interpretaciones. (van Staalduinen et al.2018). Entendiendo que el propósito de los métodos de minería de datos es extraer conocimiento significativo de datos (Han y Kamber, 2006), se puede comprender en palabras de Amerioon, S et. la (2021), que la minería de datos educativos es un campo exquisito emergente que se ha implementado con éxito en educación, él en su investigación utilizó un modelo predictivo enfocado en datos supervisados. El método propuesto tenía como objetivo demostrar por medio de la minería de datos, el descubrimiento de nuevos patrones latentes para el análisis predictivo y lograr identificar el éxito o el fracaso de los estudiantes en la Universidad de Tecnología de Shahrood.

Similares al anterior, hay diversos estudios a considerar. Por ejemplo, en ocasiones se determinan factores predictores como la demografía de los estudiantes, calificación al ingreso, puntajes en pruebas de aptitud, desempeño en cursos de primer año y su desempeño general en su programa usando la técnica de regresión, este fue el caso de Golding y Donaldson, (2006), en su estudio basado sobre los datos de una sola cohorte compuesta por 85 estudiantes de la Facultad de Informática y Tecnología de la Información en la Universidad de Tecnología de Jamaica (UTECH), encontraron una fuerte correlación entre el desempeño en un curso de ciencias de la computación

de primer año y el desempeño general de los estudiantes en el programa, con una correlación de 0.499 que explica 70,6% del rendimiento global de los alumnos.

En Asia, los investigadores Nghe, Janecek y Haddawy (2007) desarrollaron técnicas de minería de datos para predecir la rendimiento académico considerando los datos de dos institutos académicos diferentes; asiático Instituto de Tecnología (AIT), Tailandia y Universidad Can Tho (CTU), Vietnam. En esa investigación, la precisión de las predicciones se midió mediante una validación cruzada de 10 veces: 9/10 de los datos donde se utilizó para construir el modelo que se probó en 1/10 de los datos, y este proceso se repitió 10 veces. Por lo tanto, se utilizó una sola cohorte para construir el modelo de predicción y evaluarlo.

Otros estudios como el de Li, C.,et. al. (2021), está relacionado con los resultados de aprendizaje, incorporándose a modelos predictivos. Donde se tiene como objetivo de estudio explorar la posibilidad de construir un sistema de alerta temprana donde se logre predecir los resultados del aprendizaje de los estudiantes en una plataforma de aprendizaje. El modelo fue desarrollado para predecir si es probable que un estudiante apruebe o suspenda una evaluación del curso, en el estudio se usa un algoritmo seldoniano mediante el cual se logra identificar un rendimiento tanto deseablemente justo como predictivo, según los investigadores.

Otra investigación, que integra el concepto de resultados de aprendizaje en modelos predictivos es la de Othman, W., et. al (2020), donde se desarrolla un modelo predictivo basado en datos de resultados de aprendizaje del programa (PLO), en este estudio, hay dos fuentes de datos utilizadas en este estudio, la base de datos académica institucional y los comentarios en línea de los graduados, donde mediante la regresión lineal simple se logra medir el grado de relación entre la categoría de PLO y la duración del graduado para conseguir un empleo. Los resultados lograron identificar que el modelo tiene un valor potencial para ser utilizado en predecir el desempeño de la empleabilidad de los graduados dentro del marco de tiempo (6 meses) según lo determinado por el Ministerio de Educación Superior.

Referencias

Agudelo, A.S., Figueroa, L. A., & Vásquez, L. (2019). Relaciones causales de los factores que afectan el desempeño de los estudiantes en pruebas estandarizadas en Colombia. *Revista espacios, 40*(23). 1–11.

Amerioon, S., Hosseini, M. M., & Moradi, M. (2021). Extract hidden patterns in students' academic information to improve the curriculum by using data mining. International Review of Applied Sciences and Engineering, 12(3), 269-277.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer, New York, NY.

Bogoya, J. D., Bogoya, J. M., & Peñuela, A. J. (2017). Value-added in higher education: ordinary least squares and quantile regression for a Colombian case. *Ingeniería e Investigación*, 30-36.

Cifuentes, J. E., Chacón, J. A., & Fonseca, L. Á. (2019). Pruebas estandarizadas y sus resultados en una licenciatura. *Revista UNIMAR, 37*(1), 69–81. https://doi.org/10.31948/rev.unimar/unimar37-1-art4

Escobedo, M. T., Hernández, J. A., Estebané, V., & Martínez, G. (2016). Modelos de ecuaciones estructurales: Características, fases, construcción, aplicación y resultados. *Ciencia & trabajo, 18*(55), 16-22.

Feng, Z., & Wei, W. (2019, July). Study on Cultivating Students' Critical Thinking Ability Through Higher Order Questioning. *4th International Conference on Contemporary Education. Advances in Social Science, Education and Humanities Research*, (pp. 329, 759 (Vol. 762).).

García-González, J. D., & Skrita, A. (2019). Predicting academic performance based on students' family environment: Evidence for Colombia using classification trees. *Psychology, Society and Education 11*(3), 299–311. https://doi.org/10.25115/psye.v11i3.2056

García, J. R. (2020). La evaluación para el ingreso al servicio educativo y su impacto en el rendimiento de la prueba PISA en México. *Revista Iberoamericana de Educación,, 84*(1), 237-263.

Golding, P., & Donaldson, O. (2006, October). Predicting academic performance. In Proceedings. Frontiers in Education. 36th Annual Conference (pp. 21-26). IEEE.

Han, J., Kamber, M., & Mining, D. (2006). Concepts and techniques. Morgan Kaufmann, 340, 94104-3205.

Hernández-Leal, E. J., Quintero-Lorza, D. P., Escobar-Naranjo, J. C., Ramírez-Gómez, J. S., & Duque-Méndez, N. D. (2018). Educational data mining for the analysis of student desertion. Learning Analytics for Latin America 2018, 2231, 51-60.

Hyytinen, H., Toom, A. y Shavelson, RJ. (2019). Potenciar el pensamiento científico mediante el desarrollo del pensamiento crítico en la educación superior. En *Redefiniendo el pensamiento científico para la educación superior*, pp. 59-78. Palgrave Macmillan, Cham.

Irvine, J. (2017). A Comparison of Revised Bloom and Marzano's New Taxonomy of Learning. *Research in Higher Education Journal, 33.*

ICFES (2018). *¿Cómo se generan los puntajes en las pruebas saber del ICFES?.* Saber al detalle. ISSN: 2590-4663. Julio de 2018. pp. 1-6

ICFES (2019). *¿Cómo se construye el Índice de Nivel Socioeconómico (INSE) en el contexto de las pruebas saber?.* Saber al detalle. ISSN: 2590-4663. Abril de 2019. pp. 1-7.

ICFES (2021). *Medición de los efectos de la educación superior en Colombia sobre el aprendizaje estudiantil*. Bogotá: ICFES, pp.5-6.

Karlgren, K., Lakkala, M., Toom, A., Ilomäki, L., Lahti-Nuuttila, P. & Muukkonen, H. (2020). Evaluación del aprendizaje de la competencia laboral del conocimiento en la educación superior: traducción y adaptación intercultural del Cuestionario de prácticas de conocimiento colaborativo. *Artículos de investigación en educación*, *35* (1), 8-22.

Li, C., Xing, W., & Leite, W. (2021). Yet Another Predictive Model? Fair Predictions of Students' Learning Outcomes in an Online Math Learning Platform. LAK21: 11th International Learning Analytics and Knowledge Conference. doi:10.1145/3448139.3448200

Lindblom-Ylänne, S., Parpala, A. y Postareff, L. (2019). ¿Qué constituye el enfoque superficial del aprendizaje a la luz de nuevas pruebas empíricas? . *Estudios de educación superior*, *44* (12), 2183-2195.

Martínez, R., Gaviria, J. L. E., & Castro, M. (2008). Concepto y evolución de los modelos de valor añadido en educación. *Revista de educación 348.* 15-46

Martínez, A., & González, M. O. (2018). La construcción de las competencias genéricas en el nivel superior. *Revista Atlante: Cuadernos de Educación y Desarrollo.*

Melo B., L. A., Ramos F., J. E., & Hernández S., P. O. (2014). La Educación Superior en Colombia: Situación Actual y Análisis de Eficiencias. *Borradores de Economía*, 1-52.

Ministerio de Educación Nacional. (2019). *Decreto 1330 de 2019.* Bogotá: República de Colombia.

Niebles, W., Martínez-Bustos, P., & Niebles-Núñez, L. (2019). Competencias matemáticas como factor de éxito en la prueba pro en universidades de Barranquilla, Colombia. *Educación y Humanismo, 22*(38), 1–16. https://doi.org/10.17081/eduhum.22.38.3590

Nghe, N. T., Janecek, P., & Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. In 2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports (pp. T2G-7). IEEE.

OCDE. (2012). *Assessment of Higher Education Learning Outcomes Feasibility Study Report.* París: OECD.

Othman, W. N. A. W., Abdullah, A., & Romli, A. (2020, February). Predicting Graduate Employability based on Program Learning Outcomes. In IOP Conference Series: Materials Science and Engineering (Vol. 769, No. 1, p. 012018). IOP Publishing.

Oviedo, A. I., & 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. https://doi.org/10.33571/rpolitec.v15n29a10

Pedraza, N. A. (2020). Satisfacción laboral y compromiso organizacional del capital humano en el desempeño en instituciones de educación superior. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 10*(20).

Pereira, Z. (2011) Los diseños de método mixto en la investigación en educación: Una experiencia concreta. *Revista Electrónica Educare, 15*(1) 15-29

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18.

Rodrigues, M. W., Isotani, S., & Zarate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. Telematics and Informatics, 35(6), 1701-1717.

Samaja, J. (2012). *Epistemología y metodología. Elementos para una teoría de la investigación científica*. Buenos Aires: Eudeba.

Shmueli, G. (2010). To Explain or to Predict? *Statist. Sci.25*(3)289-310. doi:10.1214/10-STS330.

Soltan, G. Zh., Smailova, S. S., Uvalieva, I. M. y Tomilin, A. K. (2013). Análisis de datos intelectuales en las tareas de gestión de la calidad del proceso educativo. Educación en ingeniería, (13), 36-43.

Torres, V. G. L., Moreno, L. R. M., & Rivas, D. A. P. (2021). La formación de profesionales de las ciencias administrativas, competencias y habilidades para I4. 0. *Un agradecimiento muy especial al Instituto Politécnico Nacional (IPN) y al Centro Interdisciplinario de Investigación para el Desarrollo Integral Regional, Unidad* Oaxaca (CIIDIR Unidad Oaxaca), 290.

Van Staalduinen, De Laet, T., Broos, T., J. P., Ebner, M., & Leitner, P. (2018, April). Transferring learning dashboards to new contexts: experiences from three case studies. In Conference Proceeding Open Educational Global Conference 2018 (p. 14).

Vásquez, O. (2018). Las pruebas Saber 11 como predictor del rendimiento académico expresado en los resultados de las pruebas Saber Pro obtenidos por los estudiantes de Licenciatura en Pedagogía Infantil de la Corporación Rafael Nuñez. *Revista Científica Virtual de Pedagogía 9* (1), 187–204.

Villamil, N. E. (2019). *Aporte de la evaluación por competencias a los resultados de la organización.* http://hdl.handle.net/10654/32085.

Zapatera, A. (2021). *El método Singapur para el aprendizaje de las matemáticas [Recurso electrónico]: enfoque y concreción de un estilo de aprendizaje= The Singapore Method for the mathematics learning: approach and concretion of a learning style.* Universidad de Extremadura.

Anexo 1. Relación entre variables y su descriptor.

Descripción	Variable
Género	ESTU_GENERO
Edad	Calculado a partir de ESTU_FECHANACIMIENTO
Período de presentación del examen	PERIODO
Pertenencia a una etnia	ESTU_TIENEETNIA
Variables de discapacidad	ESTU_LIMITA_MOTRIZ ESTU_LIMITA_INVIDENTE ESTU_LIMITA_SORDO
Departamento de residencia	ESTU_DEPTO_RESIDE
Estrato socioeconómico	FAMI_ESTRATOVIVIENDA

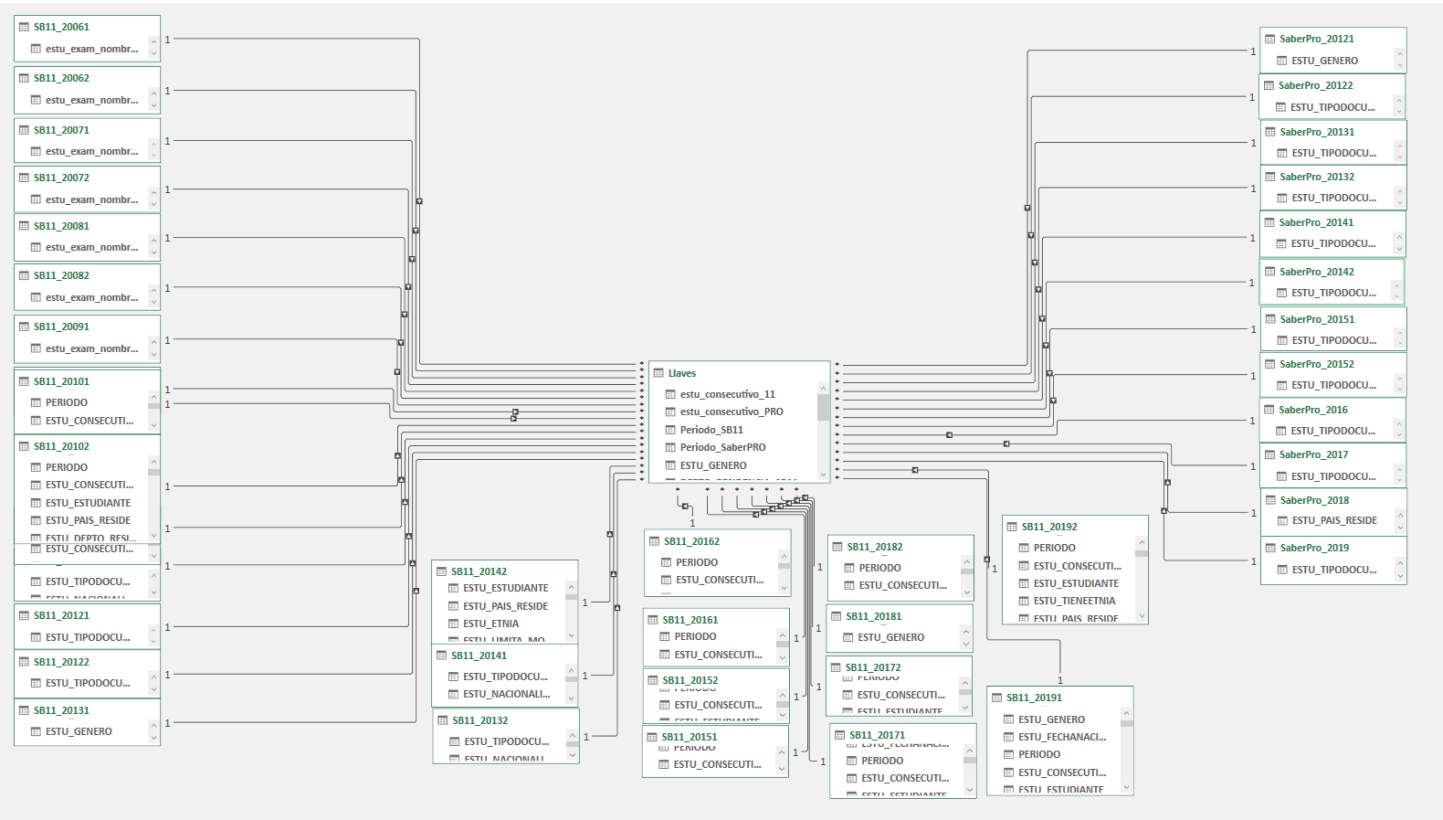
Ingreso familiar mensual	FAMI_INGRESOFMILIARMENSUAL
Situación económica	FAMI_SITUACIONECONOMICA
Género del colegio	COLE_GENERO
Naturaleza del colegio	COLE_NATURALEZA
Carácter del establecimiento	COLE_CARACTER
Puntaje en lectura crítica	PUNT_LECTURA_CRITICA
Puntaje en matemáticas	PUNT_MATEMATICAS
Puntaje en ciencias naturales	PUNT_C_NATURALES
Puntaje en sociales y ciudadanas	PUNT_SOCIALES_CIUDADANAS
Puntaje inglés	PUNT_INGLES
Puntaje total obtenido	PUNT_GLOBAL
Índice socioeconómico	ESTU_INSE_INDIVIDUAL

Relacionadas con las bases de datos de las pruebas saber PRO:

Descripción	Variable
Género	ESTU_GENERO
Edad	Calculado a partir de ESTU_FECHANACIMIENTO
Período de aplicación de la prueba	PERIODO
Estado civil	ESTU_ESTADOCIVIL
Pertenencia a una etnia	ESTU_TIENEETNIA
Variables de discapacidad	ESTU_LIMITA_MOTRIZ ESTU_LIMITA_INVIDENTE ESTU_LIMITA_SORDO
Departamento de residencia	ESTU_DEPTO_RESIDE
Estrato socioeconómico de la vivienda	FAMI_ESTRATOVIVIENDA
Ingreso familiar mensual	FAMI_INGRESOFMILIARMENSUAL
Grupo de referencia al que pertenece el programa académico del estudiante	FAMI_INGRESO_FMILIAR_MENSUAL
Ingresos mensuales del hogar	FAMI_INGRESO_FMILIAR_MENSUAL
Grupo de referencia al que pertenece el	GRUPOREFERENCIA

programa académico del estudiante	
Departamento donde se ofrece el programa	ESTU_PRGM_DEPARTAMENTO
Metodología del programa académico	ESTU_METODO_PRGM
Carácter académico de la IES	INS_CARACTER ACADEMICO
Puntaje razonamiento cuantitativo	MOD_RAZONA_CUANTITAT_PUNT
Puntaje lectura crítica	MOD_LECTURA_CRITICA_PUNT
Puntaje competencias ciudadanas	(MOD_COMPETEN_CIUDADA_PUNT
Puntaje de inglés	MOD_INGLÉS_PUNT
Nivel de desempeño en inglés	MOD_INGLES_DESEM
Puntaje comunicación escrita	MOD_COMUNI_ESCRITA_PUNT
Desempeño comunicación escrita	MOD_COMUNI_ESCRITA_DESEM).

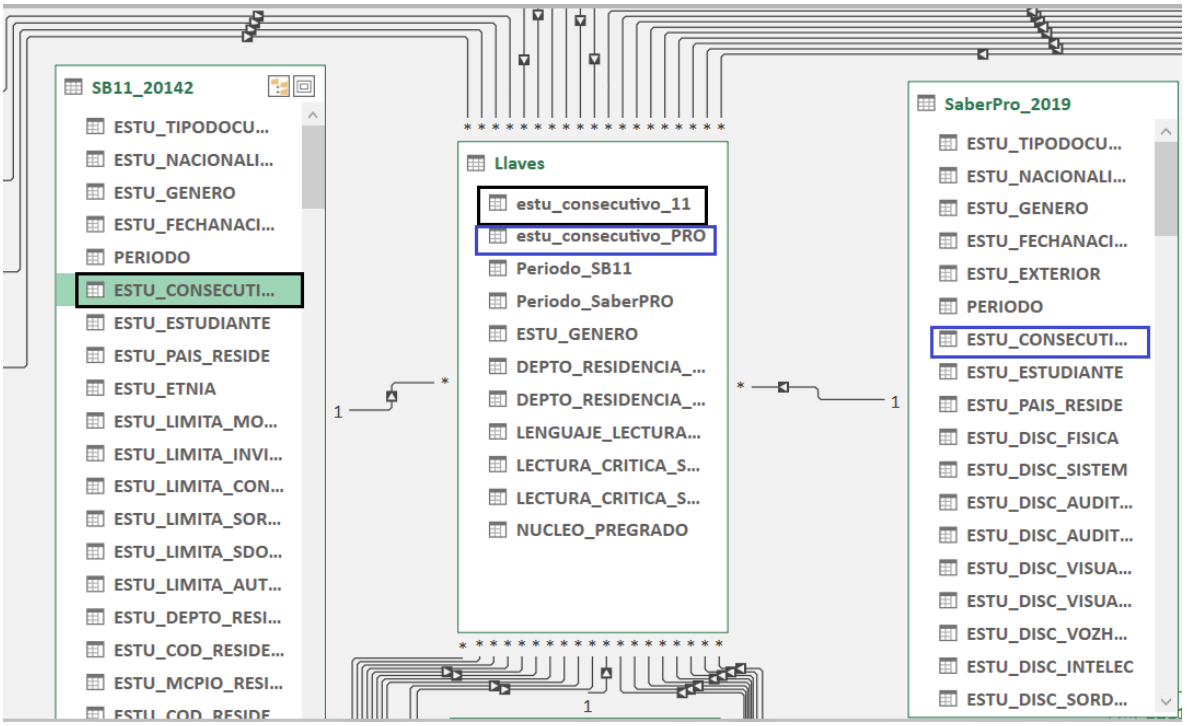
Anexo 2. Modelo de base de datos para la integración de resultados de las pruebas y las variables socioeconómicas para los períodos de estudio.



Anexo 3. Gráficos donde se presentan las asociaciones entre las bases de datos

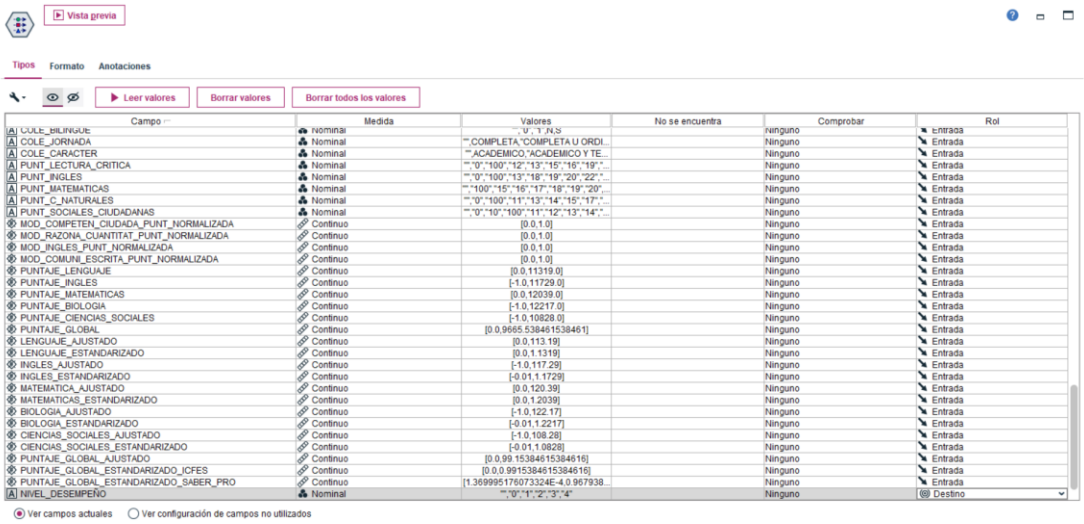
	estu_consecutiv...	estu_consecutivo_...
1	SABER1120082416622	EK201210000523
2	SABER1120082072880	EK201210000528
3	SABER1120062195960	EK201210000532
4	SABER1120072075157	EK201210000533
5	SB11201020158376	EK201210000535
6	SABER1120072031620	EK201210000536
7	SABER1120062431326	EK201210000541
8	SABER1120072214474	EK201210000543
9	SABER1120072497630	EK201210000547
10	SABER1120091052928	EK201210000549
11	SABER1120072297325	EK201210000550
12	SABER1120082050742	EK201210000553
13	SABER1120062197160	EK201210000561
14	SABER1120072480874	EK201210000562
15	SB11201120058989	EK201210000566

Como se puede observar, se generan relaciones entre los consecutivos para las pruebas saber 11 y las pruebas saber Pro. A continuación se presenta la forma en que se relacionan.



Utilizando las llaves que relacionan las pruebas Saber 11 y las pruebas Saber Pro, es posible generar consultas en las respectivas bases de datos que permiten traer los datos para su respectivo análisis, lo cual puede observarse en la siguiente figura.

Anexo 4. Campos utilizados para el cribado de variables predictoras.



Se identifican cuáles son las variables de entrada y cuál es la variable objetivo, al identificar las variables predictoras, es más factible desarrollar el modelo predictivo.