INVESTIGACIÓN e INNOVACIÓN en INGENIERÍAS
ISSN2344-8652

Modelo Estadístico para determinar los factores académicos en los Resultados de las Pruebas Saber Pro

Statistical Model to determine the academic factors in the Results of the Saber Pro Tests

Alberto Fabio Narváez Zúñiga
Universidad Del Sinú, Colombia

OPEN ACCESS

Recibido: 14/09/2022
Aceptado: 31/10/2022
Publicado: 02/12/2022

Correspondencia de autores:
albertofabionarvaez@gmail.com

Resumen

Objetivo: Diseñar un modelo estadístico que determine los factores académicos sobre resultados de las Pruebas Saber Pro. **Metodología:** El estudio de técnicas de relaciones multivariables y de aprendizaje fueron empleadas para establecer un mecanismo de relación entre un conjunto de variables académicas y sociodemográficas y su influencia con el resultado de la prueba Saber Pro, a través de un diseño y selección de un modelo estadístico multivariable que determine en forma óptima los factores académicos que inciden en los resultados de las pruebas Saber Pro. **Resultados:** Se apreció que no existen diferencias significativas entre los modelos y la realidad reflejada en la muestra de la validación a excepción de la prueba PCME que el modelo de Random Forest no prueba hipótesis de validación. Se identificó que el modelo de regresión lineal multivariante no muestra diferencias significativas en ninguna de las pruebas, al contrario del modelo Random Forest si muestra diferencias para ciertos valores de α en ING y FPI además de rechazar hipótesis de igualdad para la prueba PCME. **Conclusiones:** Cualquiera de las técnicas utilizadas en el estudio puede ayudar a realizar un modelo predictivo que sea capaz de permitir a la institución generar estrategias para lograr crear políticas orientadas a mejorar el rendimiento de los estudiantes. Sin embargo, la técnica de regresión lineal multivariante de acuerdo a las pruebas de hipótesis es la mejor posicionada en este estudio.

Palabras clave: Factores Académicos, Pruebas Saber Pro, Multivariable, Regresión Lineal, arboles de decisiones, Random Forest.

Abstract

Objective: TDesign a statistical model that determines the academic factors on the results of the Saber Pro Tests. **Methodology:** The study of multivariable relationship and learning techniques were used to establish a relationship mechanism between a set of academic and sociodemographic variables and their influence on the result of the Saber Pro test, through a design and selection of a multivariable statistical model that optimally determines the academic factors that affect the results of the Saber Pro tests. **Results:** It was observed that there are no significant differences between the models and the reality reflected in the validation sample, except for the PCME test that the Random Forest model does not test the validation hypothesis. It was identified that the multivariate linear regression model does not show significant differences in any of the tests, unlike the Random Forest model if it shows differences for certain values of α in ING and FPI, in addition to rejecting the hypothesis of equality for the PCME test. **Conclusions:** Any of the techniques used in the study can help to make a predictive model that is capable of allowing the institution to generate strategies to create policies aimed at improving student performance. However, the multivariate linear regression technique according to the hypothesis tests is the best positioned in this study.

Keywords: Academic factors, Saber Pro tests, Multivariable, Linear Regression, decision trees, Random Forest.

Como citar (IEEE): A. Narváez-Zúñiga. "Modelo Estadístico para determinar los factores académicos en los Resultados de las Pruebas Saber Pro". *Investigación e Innovación en Ingenierías*, vol. 11, n°1, 3-21, 2023. DOI: https://doi.org/10.17081/invinno.11.1.6255

Introducción

El uso adecuado de los resultados de las evaluaciones de Estado como las Pruebas Saber Pro, son un elemento indispensable para fomentar procesos de autoevaluación y mejoramiento continuo hacia el logro de altos niveles de calidad en la educación superior, razón por la cual se hace necesario contar con un compendio amplio de datos (resultados de las pruebas saber dependen de múltiples variables de carácter sociodemográfico, algunas cuantitativas, otras Dummies, etc.) e información que permitan conocer a fondo el estado de los programas académicos, sus fortalezas, debilidades y diferencias. A partir de ello, diseñar estrategias que propendan por el cierre de brechas académicas que permitan elevar el nivel de enseñanza de ciertas disciplinas a nivel general [1], sin embargo, el conjunto de datos del historial académico, junto con los resultados de la prueba de estado Saber Pro en Colombia, constituyen un gran volumen de información, con una estructura compleja, lo cual dificulta las tareas de análisis y extracción de conocimiento oculto.

Debido a esto, la minería de datos surge como una posible respuesta, a la necesidad de analizar, manipular y extraer conocimiento de los datos, toda vez que esta hace uso de sofisticados algoritmos que permiten encontrar patrones ocultos en un conjunto de datos y predecir comportamientos; por lo que el uso de la minería de datos para el análisis de información de índole académica, resulta pertinentes según estudios adelantados [2, 3, 4, 5, 6].

En estos últimos años se han adelantado iniciativas que buscan contribuir de manera efectiva al mejoramiento de los resultados obtenidos en las pruebas adelantadas a la fecha. El análisis estadístico multivariable y la selección adecuada de un modelo estadístico, se constituyen en una excelente alternativa (rigurosa y confiable) para el descubrimiento de patrones e información relevante proveniente de las bases de datos de la prueba Saber Pro del Instituto Colombiano para la evaluación de la Educación (ICFES); los resultados del presente proyecto definirán líneas de base confiables para el establecimiento de estrategias y planes de mejora para el cambio de esta realidad (deficiencia resultado pruebas saber pro) en la Escuela de Ingeniería Industrial de la universidad del Sinú, seccional Cartagena.

Se ha realizado una revisión de los antecedentes del área objeto de estudio que tienen relación con el área de uso de modelos predictivos en el contexto de rendimiento estudiantil [7], en los cuales utilizaron la definición del árbol de ascendencia y genealogía como herramientas probadas para determinar el linaje de cualquier persona y establecer dependencias entre individuos.

El árbol genealógico se puede usar aún más para obtener información sobre el investigador y su linaje escolar, que es de suma importancia en el mundo actual de la tecnología informática. Esta comprensión de la genealogía académica podría ser una forma de ayudar a los estudiantes a lograr la socialización académica dentro de la disciplina, haciendo conexiones explícitas que pueden ser influyentes. El conocimiento de su herencia científica le da al usuario una perspectiva más amplia de su propio proyecto de investigación [7], donde destacan e investigan cómo esta red académica es explotada por ciertos investigadores utilizando diversas herramientas de visualización, identificando que el factor de credibilidad e influencia está determinado por las diversas citas obtenidas por un autor y para mejorar sus rankings en diversas formas, tienden a colaborar en su círculo académico y potenciar su recuento de citas [7].

Una tendencia reciente entre los investigadores es formar comunidades basadas en sus relaciones académicas y depender de numerosas citas para su beneficio mutuo. El rastreo de las relaciones genealógicas puede ser útil para detectar dichas comunidades y también crear una métrica más consciente de la calidad utilizando un modelo independiente de linaje para el cálculo de métricas a nivel de autor.

Asimismo, según conclusión presentada en la investigación [8], analizan un gran conjunto de datos que contienen todos los cursos tomados por cada estudiante de pregrado en una importante Universidad de Canadá durante 10 años. Los algoritmos modernos de aprendizaje automático pueden utilizar grandes conjuntos de datos para crear herramientas útiles para el proveedor de datos, en este caso, la universidad.

Por otro lado, se diseña una metodología utilizada para construir reglas estructuradas en árbol. A diferencia de muchos otros procedimientos estadísticos, que pasaron del lápiz y el papel a las calculadoras, el uso de árboles en este texto era impensable antes de las computadoras. Tanto el lado práctico como el teórico han sido desarrollados en el estudio de los autores sobre métodos de árboles. Los árboles de clasificación y regresión reflejan estos dos lados, abarcando el uso de árboles como método de análisis de datos y, en un marco más matemático, demostrando algunas de sus propiedades fundamentales [9].

Un árbol de clasificación es un clasificador intuitivo y poderoso y la construcción de un bosque aleatorio de árboles mejora este clasificador. Los bosques aleatorios también permiten mediciones confiables de las variables de importancia. Estas medidas explican qué variables son útiles para los clasificadores y pueden usarse para comprender mejor qué está estadísticamente relacionado con la situación de los estudiantes. Uno de los principales objetivos de las instituciones de educación superior es brindar una educación de alta calidad a sus estudiantes y reducir las tasas de deserción. Esto se puede lograr mediante la predicción temprana del rendimiento académico de los estudiantes mediante la minería de datos educativos (EDM), teniendo como objetivo predecir las calificaciones finales de los estudiantes e identificar a los estudiantes honorarios en una etapa temprana [10].

Mejorar el rendimiento académico del estudiante no es una tarea fácil para la comunidad académica de enseñanza en nivel superior. El rendimiento académico de los estudiantes de ingeniería y ciencias durante su primer año en la universidad es un punto de inflexión en su trayectoria educativa y compromete su promedio General de puntos (GPA), del Inglés Grade Point Average, en una manera decisiva.

La aplicación de métodos de minería de datos en el campo educativo ha ganado mucha atención entre los científicos en los últimos años. La minería de datos educativos forma un área de investigación en constante desarrollo que tiene como objetivo revelar el conocimiento oculto en los datos educativos y mejorar el comportamiento y los resultados de aprendizaje de los estudiantes. Con este fin, ya se han implementado métodos de minería de datos en entornos educativos resolviendo una variedad de tareas, entre las que también se encuentra la predicción del rendimiento académico de los estudiantes.

Los árboles de decisión han demostrado ser un método bastante eficaz para problemas de clasificación y regresión, mostrando una serie de ventajas considerables, como la eficiencia, la simplicidad, la flexibilidad y la interpretabilidad. Además, la configuración de los valores de los parámetros a menudo tiene un impacto material en la construcción de árboles óptimos en términos de precisión y/o tamaño.

Los estudios adelantados en [11] describen el uso de la Minería de Datos Educativos (EDM) ayudan a reconocer el desempeño de los estudiantes y predecir sus logros académicos que incluyen los aspectos de éxitos y fracasos, aspectos negativos y desafíos. En los sistemas educativos, se ha recopilado una gran cantidad de datos de los estudiantes, lo que se ha vuelto difícil para los funcionarios buscar y obtener el conocimiento necesario para descubrir los desafíos que enfrentan los estudiantes y las universidades por métodos tradicionales. Por tanto, el problema de raíz es cómo sumergirse en estos datos y descubrir los retos reales a los que se enfrentan tanto los estudiantes como las universidades. Los resultados de este modelo se comparan con los resultados usando en una regresión lineal (RL) del inglés lineal regression [12].

La minería de datos educativos ayuda a las instituciones educativas a desempeñarse de manera efectiva y eficiente mediante la explotación de los datos relacionados con el rendimiento. Puede ayudar a los estudiantes para minimizar el riesgo de obtener una mala actuación, desarrollar sistemas de recomendación y alertar a los estudiantes en diferentes niveles. Es beneficioso para los estudiantes, educadores y autoridades en general. El aprendizaje profundo ha ganado impulso en varios dominios, especialmente en el procesamiento de imágenes con un gran conjunto de datos.

Sin embargo, en [13] diseñan un modelo de regresión para analizar el desempeño académico de los estudiantes utilizando deep learning. El modelo de aprendizaje profundo registra una puntuación absoluta media (MAE) de 1,61 y una pérdida de 4,7 con el valor de [kappa] =3. Mientras que el modelo de regresión lineal arroja una pérdida de 6,7 y una puntuación de MAE de 1,97. De acuerdo a los autores, el modelo de aprendizaje profundo supera al modelo de regresión lineal.

El modelo puede extenderse con éxito a otros programas para extraer y predecir el rendimiento de los alumnos. Predecir el éxito o el fracaso de un estudiante en un curso o programa es un problema que se ha abordado recientemente utilizando técnicas de minería de datos.

En [14], evalúan algunos de los algoritmos de clasificación y regresión más populares en de esta problemática. Se abordan dos problemas: predicción de aprobación/reprobación y predicción de calificación. El primero es abordado como tarea de clasificación, mientras que la última como tarea de regresión. Se entrenan modelos separados para cada curso. Los experimentos se llevan a cabo utilizando datos administrados del Universidad de Porto, con aproximadamente 700 cursos. Los algoritmos con mejores resultados en clasificación general fueron árboles de decisión y Maquinas de Vector de Soporte (SVM), mientras que en regresión eran SVM, Bosque aleatorio y AdaBoost. Sin embargo, en el ámbito de la clasificación, los algoritmos están encontrando patrones útiles, mientras que, en regresión, los modelos obtenidos no son capaces de batir una línea de base simple.

Análisis Multivariable

En análisis multivariable se va a desarrollar bajo los siguientes tres métodos:

Modelos de regresión lineal multivariable

La regresión lineal multivariable es la forma más común de análisis de regresión lineal. Como análisis predictivo, se utiliza para explicar la relación entre una variable dependiente continua y dos o más variables independientes a través de unos coeficientes que establecen ciertas correlaciones positivas o negativas con la variable dependiente. Las variables independientes pueden ser continuas o categóricas.

En el centro del análisis de regresión lineal multivariable está la tarea de ajustar una sola línea a través de un diagrama de dispersión. Más específicamente, la regresión lineal múltiple se ajusta a una línea a través de un espacio multidimensional de puntos de datos. La forma más simple tiene una variable dependiente y dos independientes. La variable dependiente también puede denominarse variable de resultado. Las variables independientes también pueden denominarse variables predictoras o regresores.

El resultado es la ecuación (1):

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn \quad (1)$$

Donde β0, β1,...βn son los coeficientes estimados por el modelo de regresión linealmultivariable.

Modelos de aprendizaje a través de árboles de decisión

El algoritmo de decisión pertenece a la familia de algoritmos de aprendizaje supervisado. A diferencia de otros algoritmos, el algoritmo del árbol de decisiones también se puede utilizar para resolver problemas de regresión y clasificación.

En [15], se argumenta por qué la lógica del método del árbol de regresión es más adecuada que el modelo lineal general para analizar conjuntos de datos educativos complejos. Además, aplicamos el algoritmo CART del Método del Árbol de Regresión y la Regresión Lineal Múltiple en un modelo con 53 predictores, tomando como resultado los puntajes de los estudiantes en lectura de la edición 2011 del Examen Nacional de Educación Media Superior (ENEM; N = 3.670.089), que es un conjunto de datos educativo complejo. Esta comparación empírica ilustra cómo el método del árbol de regresión es más adecuado que el modelo lineal general para proporcionar evidencia sobre relaciones no lineales, así como para tratar variables nominales con muchas categorías y variables ordinales.

El objetivo de usar un árbol de decisiones es crear un modelo de entrenamiento que se pueda usar para predecir la clase o el valor de la variable dependiente aprendiendo reglas de decisión simples inferidas de datos anteriores (entrenamiento).

El árbol de decisión, para predecir una etiqueta de clase para un registro, parte de la raíz del árbol que es el nodo con mayor impacto predictor. Se comparan los valores del atributo raíz con el atributo del registro. Sobre la base de la comparación, se sigue la rama correspondiente a ese valor y se puede dirigir al siguiente nodo. Cada nodo del árbol actúa como una prueba condicional para algún atributo de la variable que se está cuestionando con una decisión en este caso de tipo binaria (Si/No). Este procedimiento se repite hasta llegar a la última hoja que es el resultado final.

Modelos de aprendizaje a través de bosques aleatorios (Random Forest)

Se pueden utilizar para problemas de clasificación, predicción y regresión en (ML) del inglés Machine Learning. Se basa en el concepto de aprendizaje conjunto, que es un proceso de combinación de múltiples clasificadores para resolver un problema complejo y mejorar el rendimiento del modelo, en este caso utiliza arboles de decisión.

Como sugiere el nombre, Random Forest es un clasificador que contiene varios árboles de decisión en varios segmentos del conjunto total de datos analizados y toma el promedio para mejorar la precisión predictiva de ese conjunto de datos. En lugar de depender de un árbol de decisión, el bosque aleatorio toma la predicción de cada árbol y basándose en el análisis de la mayoría de las predicciones, predice el resultado final. Como desventaja principal de este método es que es un modelo de caja negra por lo que replicar su predictibilidad puede ser desventajoso si no se tienen las herramientas de software adecuadas.

Metodología

Este trabajo consideró un conjunto de datos obtenidos de la prueba Saber Pro, que es un examen dirigido a los estudiantes de educación superior próximos a graduarse.

El examen es realizado por el ICFES con el fin de medir la calidad de todas las universidades públicas o privadas en Colombia, ya sea que estén acreditadas o no, es un requisito obligatorio para todos los estudiantes que deseen adquirir el título profesional [16]. Además, el estudiante está en condición de hacerla solo si ha cursado mínimo el 75 % de los créditos de su carrera. La prueba consta de dos sesiones con duración de 4 horas y 40

minutos, la primera sesión consta de módulos de competencias genéricas y está conformada por 177 preguntas de opción múltiple con única respuesta y una pregunta abierta que evalúa la competencia de comunicación escrita, los cinco módulos evaluados en la primera sesión se pueden observar en la Tabla 1, en las que se evalúa las habilidades matemáticas, de comprensión, interpretación y evaluación de textos, comunicación de ideas por escrito, competencia comunicativa en lengua inglesa y habilidades para comprender el entorno social, ejercicio de la ciudadanía y participación activa en la comunidad, respectivamente.

Tabla 1. Competencias genéricas prueba Saber Pro.

COMPETENCIA	DESCRIPCION
Comunicación escrita (CE)	Competencia para comunicar ideas por escrito respecto a un tema en específico. En esta prueba se plantea una problemática, con la cual el estudiante deberá desarrollar un texto argumentativo
Razonamiento cuantitativo (RC)	Habilidades matemáticas que todo ciudadano debe tener, independientemente de su profesión u oficio, en las competencias de interpretación y representación, argumentación, formulación y ejecución en temas como álgebra, cálculo, geometría y estadística.
Lectura crítica (LC)	Habilidades para comprender, interpretar y evaluar textos, entender el significado de palabras frases, relacionar las partes de un texto para darle sentido global, determinar si las razones del autor son o no convincentes e identificar argumentos y supuestos
Competencia ciudadana (CC)	Conocimiento y habilidades necesarias para comprender el entorno social, sus problemáticas y analizar diversas posturas en situaciones de conflicto, así como competencias en la argumentación, conocimientos, multiperspectivismo y pensamiento sistémico
Inglés (ING)	Competencia comunicativa en lengua inglesa a partir de pruebas de lectura, léxico y gramática

Fuente: Elaboración propia.

En la segunda sesión son evaluadas las competencias específicas, estas constan de alrededor de 140 preguntas e incluyen módulos referentes a las áreas propias de la carrera [16], en la Tabla 2 se mencionan los módulos evaluados para la ingeniería industrial.

A partir de la base de datos del ICFES se filtró por universidad, carrera, año, sin modificar las variables reportadas y se seleccionan 158 observaciones de estudiantes que presentaron la prueba SABER PRO durante los años de 2016 a 2019.

Las variables seleccionadas de la base de datos de ICFES se muestran en la Tabla 3, a excepción de la variable CLASE que se procuró con la información que fue entregada por la universidad.

Tabla 2. Competencias específicas prueba Saber Pro.

COMPETENCIA	DESCRIPCION
Formulación de proyectos de Ingeniería (FPI)	Reconocimiento e identificación de condiciones relevantes para la caracterización y formulación de proyectos. Formulación y evaluación del proyecto. Reconocimiento del papel y responsabilidad disciplinar, social y ética como ingeniero en un contexto de desempeño profesional
Diseño de sistemas productivos y logísticos (DSPL)	Producción de bienes y servicios. Logística, métodos cuantitativos
Pensamiento científico matemáticas y estadística (PCME)	Matemáticas, estadística.

Fuente: Elaboración propia.

Tabla 3. Variables seleccionadas prueba Saber Pro.

VARIABLE	DESCRIPCIÓN	TIPO
GENERO	El género del estudiante	Cualitativa
MBECA	El estudiante paga de la matricula mediante beca	Cualitativa
MCREDITO	El estudiante paga de la matricula mediante crédito	Cualitativa
MPADRES	El estudiante paga de la matricula mediante los padres	Cualitativa
MPROPIO	El estudiante paga de la matricula medios propios	Cualitativa
EPREPARA	El estudiante tomó preparación para el examen	Cualitativa
EPADRE	La educación del padre del estudiante	Cualitativa
EMADRE	La educación de la madre del estudiante	Cualitativa
ESTRATO	El estrato socioeconómico del estudiante	Cualitativa
INTERNET	El estudiante tiene acceso a internet	Cualitativa
TV	El estudiante tiene acceso a servicios de televisión	Cualitativa
COMPUTADOR	El estudiante tiene computador propio	Cualitativa
LAVADORA	El estudiante tiene lavadora	Cualitativa
HORNO	El estudiante tiene horno	Cualitativa
AUTO	El estudiante tiene automóvil	Cualitativa
MOTO	El estudiante tiene motocicleta	Cualitativa
TRABAJA	El estudiante trabaja	Cualitativa
CLASE	Clase a la que pertenece el estudiante según su índice académico (1) 3,595 y 3,742 (Excluido) (2) 3,742 y 3,888 (3) 3,888 y 4,035 (4) 4,035 y 4,182 (5) 4,182 y 4,328 (6) 4,328 y 4,475	Cualitativa

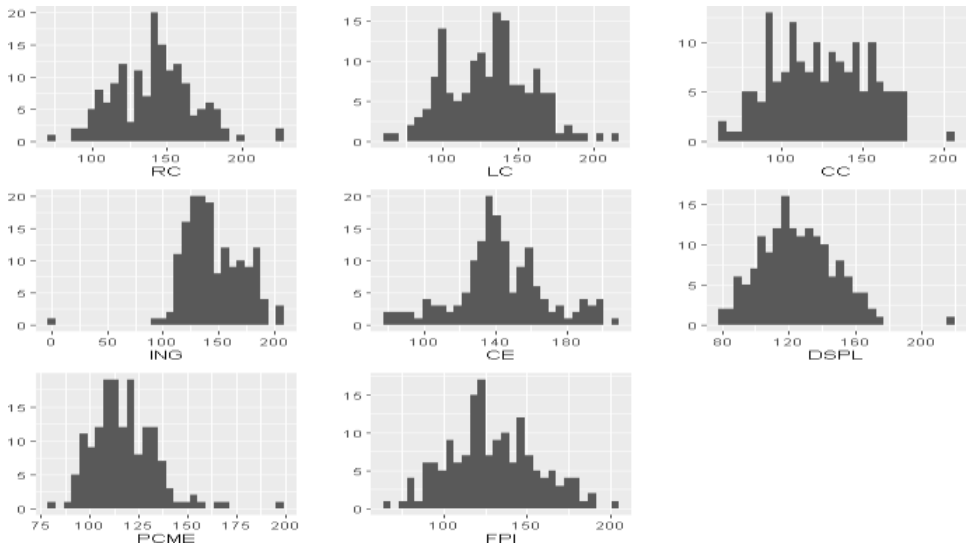
Fuente: Elaboración propia.

Metodológicamente se ha considerado el estudio de técnicas de relaciones multivariables y de aprendizaje para establecer un mecanismo de relación entre un conjunto de variables académicas y sociodemográficas y su influencia con el resultado de la prueba Saber Pro, a través de un diseño y selección de un modelo estadístico multivariable que determine en forma óptima los factores académicos que inciden en los resultados de las pruebas Saber Pro, para ello se adelantaron dos etapas, la primera etapa corresponde a un análisis previo (análisis exploratorio) de los datos sociodemográficos y promedios académicos, donde se realizó la comprobación de hipótesis de normalidad, así como la correlación y homocedasticidad entre las variables dependientes, seguido de la verificación de la fiabilidad de los datos descritos anteriormente.

Análisis descriptivo de las variables dependientes

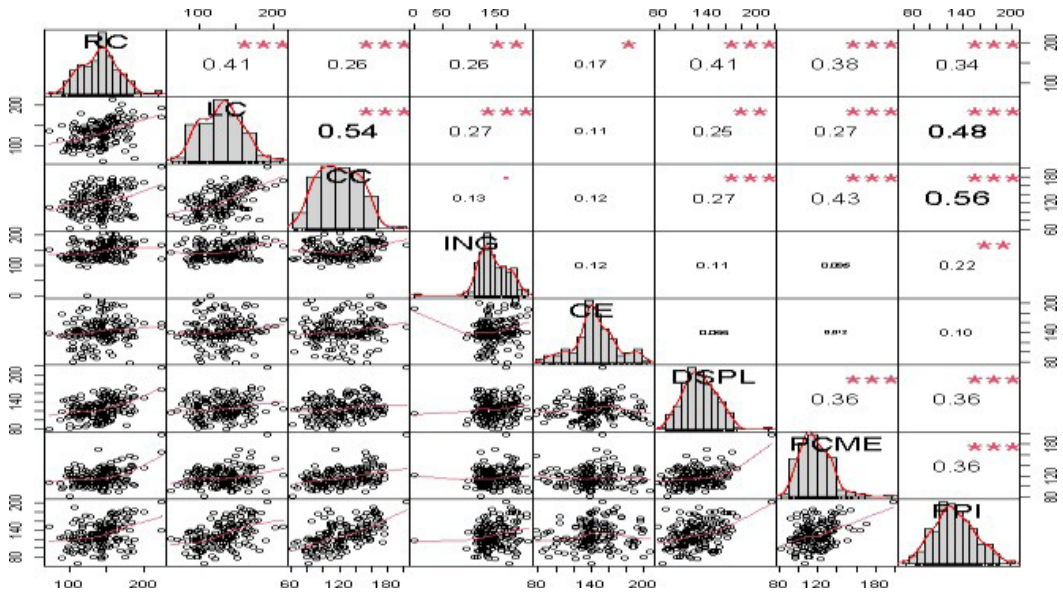
La Figura 1 presenta el histograma de frecuencia de las variables dependientes que se definen como continuas y están censuradas porque no se tiene en la muestra valores que puedan respaldar todo el rango de datos teóricos que pueden tener las variables objeto de estudio.

Figura 1. Histograma de frecuencia de las variables dependientes.



Fuente: Elaboración propia.

Figura 2. Gráfica de correlación.



Fuente: Elaboración propia.

Un análisis de simetría comprueba cierto nivel de sesgo en las pruebas PCME e ING con 1,222 y -0,697 respectivamente. Un gráfico de correlación con histogramas, funciones de densidad, líneas de regresión suavizadas y coeficientes de correlación con los niveles de significación se presenta en la Figura 2, en él se puede observar la correlación existente entre las diferentes pruebas (si no hay estrellas, la variable no es estadísticamente significativa y no existe correlación, mientras que una, dos y tres estrellas significan que la correspondiente variable es estadísticamente significativa para los niveles 10 %, 5 % y 1 %, respectivamente).

Adicionalmente se observa, una línea que pasa por la nube de puntos que da cuenta de un análisis de regresión También se observa empíricamente que la variable ING y PCME no presentan una línea recta suavizada en sus respectivas nubes de puntos correlaciónales.

Las pruebas de normalidad se utilizan en el análisis estadístico para determinar si una distribución normal es aceptable como modelo para los datos analizados. Una amplia gama de pruebas disponibles emplea diferentes propiedades de la distribución normal para comparar distribuciones empíricas y teóricas [17]. Una prueba estadística de normalidad basada en sobre la prueba de normalidad de Kolmogorov-Smirnov con media y varianza desconocidas establece que no existen diferencias significativas para probar la hipótesis nula de que los datos provienen de una población con distribución normal para todas las variables excepto CE, PCME e ING. Además de estas pruebas se desarrollaron pruebas de homocedasticidad de Levene, lo que nos permite afirmar que se cumple hipótesis de igualdad de varianzas en todas las variables dependientes.

Se han "linealizado" todas las variables independientes presentadas en la Tabla 4 al hacerlas" Dummies", con un posible valor dicotómico (0,1) lo que permite garantizar cierta relación de linealidad entre la variable dependiente y las independientes.

Tabla 4. Descripción Variables académicas y sociodemográficas prueba Saber Pro.

VARIABLE	NIVELES	RESPUESTA
GENERO	2	F/M
MBECA	2	SI/NO
MCREDITO	2	SI/NO
MPADRES	2	SI/NO
MPROPIO	2	SI/NO
EPREPARA	2	SI/NO
EPADRE	12	No Aplica / No sabe / Ninguno / Primaria incompleta / Primaria completa / Secundaria (Bachillerato) incompleta / Secundaria (Bachillerato) completa / Técnica o tecnológica incompleta / Técnica o tecnológica completa / Educación profesional incompleta / Educación profesional completa / Postgrado
EMADRE	12	No Aplica / No sabe / Ninguno / Primaria incompleta / Primaria completa / Secundaria (Bachillerato) incompleta / Secundaria (Bachillerato) completa / Técnica o tecnológica incompleta / Técnica o tecnológica completa / Educación profesional incompleta / Educación profesional completa / Postgrado
ESTRATO	6	Estrato 1 / Estrato 2 / Estrato 3 / Estrato 4 / Estrato 5 / Estrato 6
INTENET	2	SI/NO
TV	2	SI/NO
COMPUTADOR	2	SI/NO
LAVDORA	2	SI/NO
HORNO	2	SI/NO
AUTO	2	SI/NO
MOTO	2	SI/NO
TRABAJA	5	0/ Menos de 10 horas / Entre 11 y 20 horas / Entre 21 y 30 horas / Más de 30 horas
CLASE	6	1 / 2 / 3 / 4 / 5 / 6

Fuente: Elaboración propia.

La segunda etapa corresponde a un análisis y modelado multivariable en la que se emplean diferentes métodos (Análisis estadísticos, Análisis de Clúster, Métodos de Regresión Múltiple, arboles de decisión, bosques aleatorios, entre otros), con el fin de seleccionar el modelo de predicción óptimo, aquel que permita determinar, las variables independientes más significativas, así como su relación e influencia.

Resultados y Discusiones

Los experimentos se han desarrollado con el análisis de los datos de la muestra de 158 observaciones, para cada uno de los tres modelos mencionados en el análisis multivariable. Para esto se usaron las bondades que proporciona la plataforma R-Studio1 [18].

Para la experimentación se ha particionado la muestra en dos segmentos aleatorizados. Un segmento para datos de entrenamiento de los modelos y otro para pruebas estadísticas de validación. Para la segmentación de los datos de entrenamiento se han escogido al azar 126 muestras, el 80 % del total de la muestra. El segmento de la muestra donde se realizó la validación es de un tamaño de 32 observaciones (20 %) atendiendo a lo dispuesto en el teorema del límite central y a la inferencia estadística tomando como base normalidad de dicha muestra. Para la inferencia de la validación se ha considerado hacer prueba estadística de observaciones pareadas [19]. Los significados de las estadísticas están fuertemente vinculados a los indicadores de los diferentes niveles de razonamiento inferencial presentados aquí. Además, cada nivel está asociado con el razonamiento inferencial informal, preformal o formal. Los niveles propuestos de Razonamiento Inferencial para lat-Alumno se espera que la estadística y sus indicadores sean útiles para el diseño de actividades que promuevan gradualmente el razonamiento inferencial formal basado en el razonamiento inferencial informal sobre esta estadística [20].

Resultados Experimentales

Esta fase se dividió en dos temas, modelado y entrenamiento además de la fase de validación.

Pruebas de modelado y entrenamiento

Para cada una de las ocho variables dependientes clasificadas en competencias genéricas y especificas se ha aplicado modelamiento con los datos de entrenamiento, ver Tablas 1 y 2.

Regresión Lineal Multivariante

A continuación, se presenta resultado de la regresión lineal multivariante para cada una de las ocho pruebas.

Coeficientes de las variables de predicción que se deben tomar en cuenta para modelar la regresión lineal. En las Tablas 5 y 6 se observan los resultados de la salida presentada por el software R-Studio para la variable dependiente CC, incluido los Valores mínimo, Primer cuartil, media, tercer cuartil y máximo de los coeficientes de variables independientes. En dicha tabla se puede observar que variables son relevantes de acuerdo a la salida del modelo (aquellas que tienen un Pr(> |t|) bajo), además presenta aquellas que no lo son y las que no se toman en cuenta (NA) pues no revisten ninguna significación por cuanto en la muestra están compuestas de un solo valor.

Tabla 5. Salida del software R-Studio variable dependiente CC.

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	172.48836	39.50498	4.366	2.91e-05	***
GENERO	3.39992	5.33180	0.638	0.52504	
MBECA	-51.40438	17.58402	-2.923	0.00422	**
. . .	. . .	. . .	. . .	. . .	
EPREPARA3	NA	NA	NA	NA	
EPADRE1	23.19417	12.58439	1.843	0.06806	.
EPADRE6	40.35629	19.47626	2.072	0.04064	*
EPADRE10	21.07850	12.24848	1.721	0.08813	.
ESTRATO2	-37.80425	20.22820	-1.869	0.06435	.
ESTRATO3	-37.90742	21.24050	-1.785	0.07712	.
TRABAJA0	-13.23289	7.09447	-1.865	0.06486	.
CLASE6	20.61316	12.19622	1.690	0.09389	.
Signif.codes : ***0,001 ** 0,01 * 0,05,0,1””1					

Fuente: Elaboración propia

Tabla 6. Valores mínimo, Primer cuartil, media, tercer cuartil y máximo de los coeficientes de variables independientes.

Residuals				
Min	1Q	Median	3Q	Max
-58.478	-16.190	-0.815	15.506	58.077
Coefficients: (6 not defined because of singularities (NA))				

Fuente: Elaboración propia

La base de una regresión lineal múltiple es evaluar si una variable dependiente continua se puede predecir a partir de un conjunto de variables independientes (o predictoras). O en otras palabras, cuánta varianza en una variable dependiente continua se explica por un conjunto de predictores. Ciertos enfoques de selección de regresión son útiles para probar predictores, lo que aumenta la eficiencia del análisis.

A partir de la salida de ese modelo se pueden aplicar algunas técnicas para una mejor selección de los predictores y coeficientes del modelo de regresión lineal multivariante para minimizar la varianza residual R2 (que es la que no se puede explicar por el conjunto de variables predictoras). Se pueden utilizar básicamente tres procedimientos de selección para producir la ecuación de regresión más apropiada: selección hacia adelante (Forward Selection), selección hacia atrás (Backward Selection), selección por pasos (Stepwise Selection). Los tres procedimientos se consideran métodos de regresión estadística. Muchas veces los investigadores utilizan métodos de entrada de regresión secuencial (jerárquica o por bloques) que no se basan en resultados estadísticos para seleccionar predictores. La entrada secuencial permite al investigador un mayor control del proceso de regresión. Los elementos se ingresan en un orden determinado basado en la teoría, la lógica o la practicidad, y son apropiados cuando el investigador tiene una idea de qué predictores pueden afectar la variable dependiente.

En este caso, se ha seleccionado el método de selección por pasos (Stepwise Selection) para efectos de presentar el resultado porque es una combinación de los dos anteriormente mencionados. La selección de predictores empleando stepwise selection (hybrid/doble) ha identificado como mejor modelo el formado por los predictores, a continuación se relacionan en Tabla 7 valores de coeficientes de variables independientes para CC.

Tabla 7. Valores de los coeficientes de las variables independientes para CC.

(Intercept)	MBECA	EPREPARA2	EPADRE1	EPADRE4	EPADRE6	EPADRE8
181.659	-51.892	18.094	25.454	53.547	47.273	14.890
EPADRE9	EPADRE10	EPADRE11	EMADRE2	EMADRE3	EMADRE5	EMADRE6
15.270	21.904	19.452	-22.231	25.533	28.614	-38.988
EMADRE7	EMADRE9	EMADRE11	ESTRATO1	ESTRATO2	ESTRATO3	ESTRATO4
15.739	8.686	10.049	-23.027	-31.547	-33.267	-29.315
HORNO	TRABAJA0	TRABAJA2	CLASE2	CLASE3	CLASE4	CLASE5
7.756	-11.055	-12.597	-2.781	-6.071	2.174	4.427
CLASE6						
21.492						

Fuente: Elaboración propia

Las variables independientes son cualitativas y para efectos de la experimentación se han convertido en variables "Dummy" que son categóricas y solo toman el valor de 0 o de 1, para indicar la presencia o no de una categoría especifica. La ecuación de regresión lineal para el caso de la prueba de competencias ciudadanas (CC) quedaría de acuerdo a lo anterior y a lo descrito en la ecuación (2):

$$CC = 181{,}659 - 51{,}892*MBECA + 18{,}094*EPREPARA2 + 25{,}454*EPADRE1 + 53{,}547*EPADRE4 + 47{,}273*EPADRE6 + + 4{,}427*CLASE5 \quad (2)$$

Se puede notar cual es la correlación que tiene cada una de las variables cualitativas (+/–) y su efecto en la variable dependiente CC. Ejemplo de esto es la variable MBECA (El estudiante paga de la matricula mediante beca) que si su valor predictivo es uno (1) es decir "NO" la correlación negativa hace que según el modelo de regresión se le resten 51.892 puntos a la nota de partida que es 181.659. El impacto de MBECA es bastante importante y parece ser un factor determinante del éxito para la prueba CC. Además de estas pruebas se desarrollaron pruebas de homocedasticidad de Levene lo que permite afirmar que la predicción de los modelos descritos en esta sección para las variables dependiente producen un valor en los rangos en los que el modelo de regresión lineal multivariante puede trabajar.

Para las siete competencias restantes el modelo de regresión lineal multivariante da como resultado lo siguiente:

$$RC = 135{,}054 - 19{,}509*GENERO - 15{,}089*MBECA - 10{,}495*MPROPIO + 35{,}152*EPREPARA1 + 47{,}093 * EPREPARA2 + + 8{,}274 * TRABAJA3 \quad (3)$$

$$LC = 198{,}298 - 8{,}760*GENERO - 30{,}793*MBECA + 14{,}073*EPREPARA2 + 70{,}829*EPADRE4 + 43{,}531*EPADRE6 + + 23{,}107*CLASE6 \quad (4)$$

$$ING = 164{,}502 - 7{,}273*GENERO + 10{,}085*EPREPARA2 + 38{,}189*EPADRE4 + 24{,}024*EPADRE6 - 19{,}916*EPADRE8 + + 6{,}372*TRABAJA0 \quad (5)$$

$$CE = 148{,}236 + 23{,}085*EPREPARA1 + 32{,}724*EPREPARA2 + 20{,}528*EPADRE3 + 21{,}896*EPADRE8 + - 11{,}024*EMADRE12 \quad (6)$$

$$DSPL = 124{,}141 - 8{,}165*GENERO + 7{,}964*EPREPARA2 - 17{,}223*EMADRE2 + 31{,}966*EMADRE6 + + 7{,}729*HORNO \quad (7)$$

$$FPI = 123{,}609 - 34{,}954*MBECA - 8{,}925*MPROPIO + 20{,}615*EPREPARA1 + 42{,}798*EPREPARA2 + + 29{,}829*CLASE6 \quad (8)$$

$$PCME = 117{,}529 - 4{,}750*MCREDITO - 5{,}798*MPADRES + 4{,}689*EPADRE11 + + 3{,}963*MOTO \quad (9)$$

Arboles de decisión

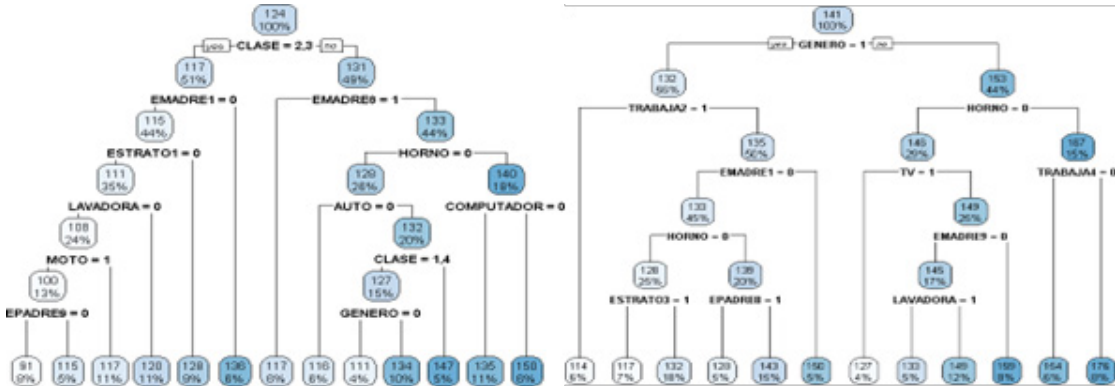
Utilizando la plataforma R-Studio se halla la mejor relación entre las variables independientes y las salidas de las variables dependientes.

Los árboles de decisión aquí construidos soportan su composición a través de nodos los cuales arrancan desde un nodo raíz que se supone es el mejor predictor. Al final se acercan a una hoja que es el valor de la estimación de la variable dependiente.

El algoritmo hace segmentación y agrupamiento sucesivo calculando una función de entropía hasta que llega a la solución final, para lo cual se ha utilizado la librería de R- studio llamada rpart.

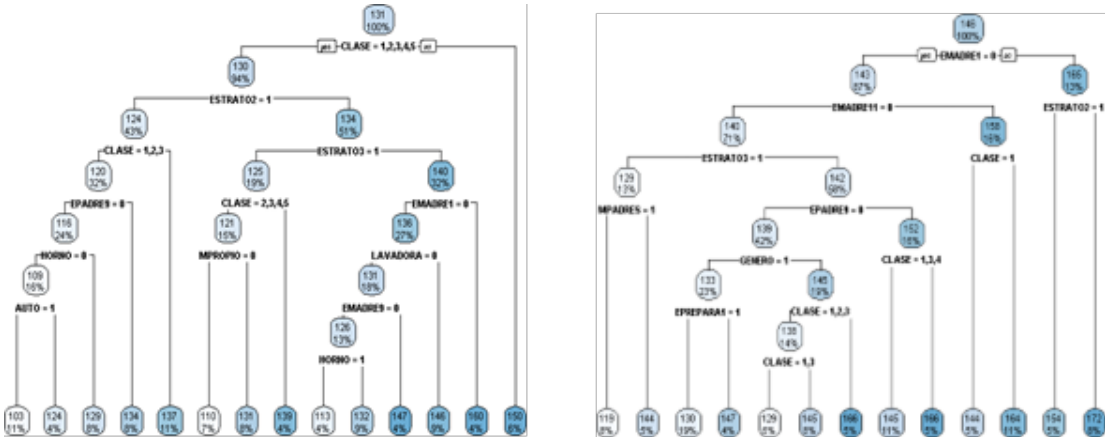
En las Figuras 3 y 4 se puede observar los árboles de decisión generados por el algoritmo. Podemos observar que en la mayoría de ellos se observa como nodo raíz la variable independiente CLASE que es el promedio de notas que trae el estudiante en su desempeño académico y universitario.

Figura 3: Arboles de decisión propuestos por rpart: CC (izquierda), RC (derecha).



Fuente: Elaboración propia.

Figura 4. Arboles de decisión propuestos por rpart: LC (izquierda), ING (derecha).



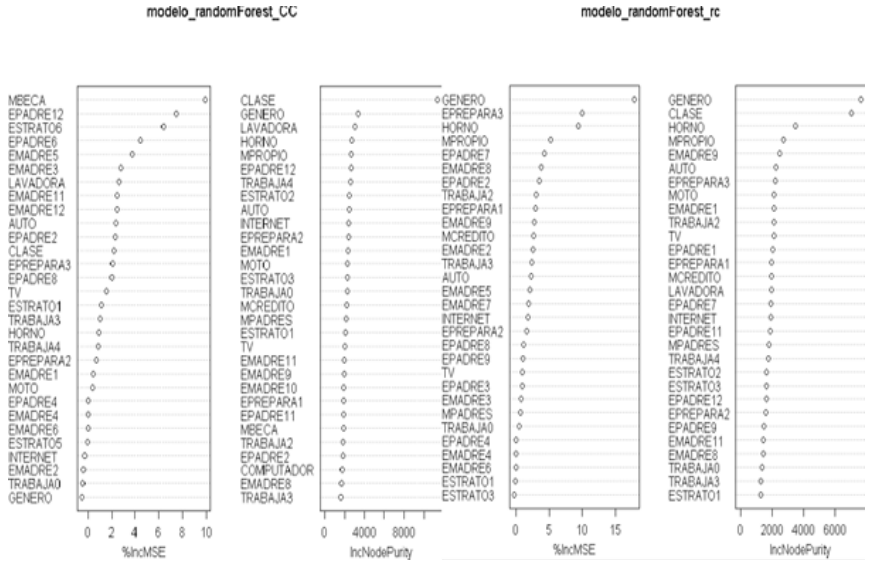
Fuente: Elaboración propia.

Modelo de bosques aleatorios (Random Forest)

Para esta experimentación se ha utilizado la librería llamada random Forest de R-Studio con valores de hiper parámetros mtry=5 (Cantidad de variables para ser modificadas aleatoriamente) y ntree=1000 (Numero de árboles). En las Figuras 5 y 6 se puede observar la importancia relativa de las variables predictoras. Para esta interpretación se enfoca el estudio de la variable gráficada como %IncMSE la cual nos da una relación de mayor a menor medida de aumento del MSE (Error Cuadrático Medio) explicada en los modelos teóricos escalados.

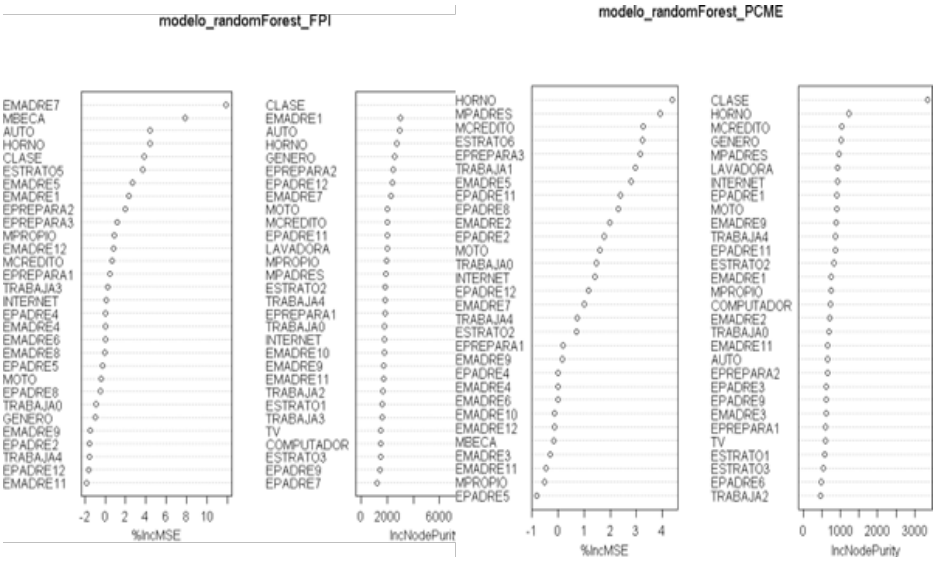
En la Figura 5 se observa que para la prueba Conciencia Ciudadana (CC) y Razonamiento cuantitativo (RC) el modelo presentado por Random Forest se parece al de regresión lineal multivariable en cuanto a que ambos escogen las mismas variables predictoras. MBECA en el caso de CC y GÉNERO en el caso de RC. La medida IncNodePurity da cuenta de aquellas variables que determinan más incremento en la pureza de los nodos. Es una medida que contiene mucho sesgo por lo que se debe usar muy poco para procesos de interpretación.

Figura 5: Variables de Importancia Random Forest: CC (izquierda), RC (derecha).



Fuente: Elaboración propia.

Figura 6. Variables de Importancia Random Forest: CE (izquierda), DSPL (de- recha).



Fuente: Elaboración propia.

Comparación de los modelos Regresión Lineal Multivariante, Arboles de Decisión y Bosques Aleatorios.

Se realiza la comparación de los tres modelos descritos. Para esto se utilizó las 32 muestras de las cuales conocemos sus valores para las pruebas que se están intentando predecir. Para la comparación se utilizó una prueba de hipótesis de diferencia de medias, H0: $\bar{X}_D = 0$ (10) llamada de observaciones pareadas (Shannon, 1988). Se ejecutan los tres modelos para la predicción y los resultados los comparamos con la realidad observada en la muestra.

En la Tabla 8 se pueden observar los resultados. En ellos se puede apreciar que no existen diferencias significativas entre los diferentes modelos y la realidad reflejada en la muestra de la validación a excepción de la prueba PCME que el modelo de Random Forest no prueba hipótesis de validación. Analizando con más detalle la tabla, se puede notar que el modelo de regresión lineal multivariante no muestra diferencias significativas en ninguna de las pruebas, al contrario del modelo Random Forest que si muestra diferencias para ciertos valores de α en ING y FPI además de rechazar hipótesis de igualdad para la prueba PCME.

Tabla 8. Resultado numérico de la prueba de observaciones pareadas. La (t∗) calculada, media muestral (x) y el error típico con respecto a los datos reales (σxy).

Hipótesis de igualdad de medias α=0,1(°) Valor Tabla=1,6955, α=0,05 (•) Valor Tabla = 2,0395, and α = 0,01 (▪) Valor Tabla = 2,7440

	Regresión lineal M				Arboles de Decisión				Bosques Aleatorios			
Prueba	t*	x	$x\pm\sigma xy$		t*	x	$x\pm\sigma xy$		t*	x	$x\pm\sigma xy$	
CC	0.1376	120.2210	120.2210±14.846	◦	0.4083	122.8720	122.8720±13.3510	◦	0.3307	122.8480	122.8480±8.020	◦
RC	0.2276	135.8580	135.8580±12.8520	◦	0.9033	139.7810	139.7810±14.6490	◦	1.4080	139.8550	139.8550±4.5620	◦
LC	1.4865	126.8421	126.8421±15.2019	◦	0.2486	129.8183	129.8183±11.9603	◦	1.0505	132.3761	132.3761±4.8468	◦
ING	1.6968	148.2431	148.2431±11.0086	•	1.3361	149.3282	149.3282±12.2939	◦	1.6967	149.6702	149.6702±3.9801	•
CE	0.2946	141.1391	141.1391±12.3671	◦	0.1323	139.7283	139.7283±11.6268	◦	0.2198	140.6837	140.6837±3.0868	◦
DSPL	0.1407	123.4306	123.4306±10.3476	◦	0.0115	123.9485	123.9485±10.6029	◦	0.4758	125.0896	125.0896±2.9578	◦
FPI	0.1959	125.7724	125.7724±16.0406	◦	0.5143	128.8229	128.8229±14.1392	◦	2.5648	135.1198	135.1198±3.8991	▪
PCME	1.5986	112.2645	112.2645±5.9481	◦	2.6867	114.5715	114.5715±5.1043	▪	4.1307	114.9622	114.9622±1.9115	

Fuente: Elaboración propia.

Los modelos propuestos pueden servir para hacer un modelo operacional de predicción y generar algunas estrategias de corrección en el estudiantado con miras a mejorar los índices de eficiencia.

Análisis de los factores independientes y sus diferentes niveles de preponderancia con la competencia en las diferentes propuestas los modelos sugieren la importancia de ciertas variables independientes en el proceso de la obtención del valor de una variable dependiente. A continuación, se enumeran las variables que pueden afectar el rendimiento en alguna competencia:

Competencia Ciudadana (CC): En competencia Ciudadana existe una variable que sobresale en los tres modelos y es señalada como un elemento predictivo que debe otorgársele importancia para la predicción. Es la variable que se define como CLASE. Teniendo impacto en diferentes direcciones dependiendo del estrato de esa clase donde se encuentre el estudiante. CLASE es el promedio del índice académico que tiene el estudiante al momento de realizar la prueba Saber Pro. De acuerdo a los modelos, aquellos que tienen índices académicos altos (que pertenezcan a la clase 4, 5 y 6) son estudiantes con una tendencia a obtener un resultado positivo que incrementa el nivel base de la variable dependiente.

Razonamiento Cuantitativo (RC): En esta competencia los modelos le dan una preponderancia especial al factor independiente GÉNERO. El análisis inicial sugiere a GENERO como una característica especial y de una correlación positiva para el caso en que toma un valor de 1 (FEMENINO) y negativo en caso contrario (MASCULINO). Sugiere generar alguna estrategia de refuerzo para revertir en el caso masculino esta correlación.

Lectura crítica (LC): Los estudiantes que tienen índice académico identificado como CLASE 6 (Promedio índice académico > 4,328) tienen una correlación positiva preponderante con la variable dependiente.

Inglés (ING): El papel de la madre y el padre con su base educativa parece jugar un rol fundamental para la prueba de inglés. Así las variables identificadas como

EMADRE, EPADRE con educación completa a nivel técnico o profesional demuestran una correlación positiva con la prueba de inglés del estudiante.

Comunicación escrita (CE): CLASE, EMADRE Y EPADRE son las variables independientes con mayor preponderancia y que generan un gran impacto en la variable dependiente CE.

Diseño de sistemas productivos y logísticos (DSPL): GENERO con correlación negativa y la propiedad sociodemográfica HORNO son las variables independientes con mayor preponderancia y que generan un gran impacto en la variable dependiente CE.

Formulación de proyectos de ingeniería (FPI): CLASE, EMADRE Y EPADRE son las variables independientes con mayor preponderancia y que generan un gran impacto en la variable dependiente FPI.

Pensamiento científico matemáticas y estadística (PCME): HORNO con correlación positiva, EMADRE Y EPADRE son las variables independientes con mayor preponderancia y que generan un gran impacto en la variable dependiente PCME.

Este estudio establece algunas variables preponderantes donde se revisó con detenimiento y propiciar políticas y estrategias de refuerzo. En la Figura 7 (izquierda) se puede apreciar una nube con las variables independientes que más aparecen en los tres modelos descritos en secciones precedentes. El tamaño indica la frecuencia de aparición.

Figura 7. Variables de Importancia General (izquierda), Especifica (derecha).

Fuente: Elaboración propia.

Se puede observar que aparecen de manera consistente y en mayor preponderancia las variables EPADRE, EMADRE, HORNO y CLASE. Por otra parte, en la Figura 7 (derecha) se pueden observar con más detalle las variables independientes y en el caso de variables con diferentes significados aparece al lado de estas la opción que más frecuencia tiene de acuerdo a lo descrito en la tabla 4.

Conclusiones

En este artículo se ha realizado el estudio de diferentes técnicas para el modelaje del resultado de las pruebas Saber-Pro. Dichas técnicas presentan una variedad de alternativas que nos permiten abarcar un amplio margen de investigación y desarrollar modelos operacionales para actuar proactivamente en la consecución del objetivo institucional que es el de mejorar el rendimiento académico de los estudiantes de universidades.

Se realiza experimentación y comparación de tres técnicas de modelamiento como lo son: Modelos de regresión lineal multivariable, Modelos de aprendizaje a través de árboles de decisión, Modelos de aprendizaje a través de bosques aleatorios. Con un marco de datos conformado por un grupo diverso de variables de tipo académica y sociodemográficas en una temporalidad delimitada durante los años 2016 al 2019.

Del estudio se puede inferir que cualquiera de las técnicas mencionadas anteriormente puede ayudar a realizar un modelo predictivo que sea capaz de permitir a la institución el generar estrategias para lograr crear políticas orientadas a mejorar el rendimiento de los estudiantes. Sin embargo, la técnica de regresión lineal multivariante de acuerdo a las pruebas de hipótesis es la mejor posicionada en este estudio para cumplir el objetivo de la predictibilidad. El modelo demostró ser capaz de lograr buenos márgenes de predictibilidad al ser validado con los datos definidos para tal fin. Esta técnica es sencilla de implementar en un marco operacional.

Aunado a esto se puede observar las relaciones principales entre las variables dependientes del modelo y las independientes. Esto permite en consecuencia determinar aquellos factores promotores del éxito en la presentación de la Prueba Saber Pro. Una de los factores preponderantes y que tiene presencia en la mayoría de las pruebas es la variable independiente CLASE que describe el índice de notas promedio de un estudiante en el momento de presentar la prueba. La formación y educación de los padres también es un factor fundamental para lograr un buen rendimiento en las pruebas de competencias genéricas. Mención aparte merece la variable GENERO como un factor que denota una influencia en algunas pruebas de competencia genérica y especifica.

Esta investigación introduce y aplica las técnicas de minería de datos para la toma de decisiones en la institución lo que permite generar estrategias de prevención y adecuación de políticas para el mejoramiento de los índices y rendimiento de nuestros estudiantes.

Referencias bibliográficas

1. Consejo Nacional de Acreditación (CNA). *Lineamientos y aspectos por evaluar para la acreditación en alta calidad de programas académicos*.[en línea].2017. Disponible en https://www.cna.gov.co/1779/articles-404750_norma.pdf.
2. J. Ruby y K. David, K. An Analysis on Academic Performance of Students using a Hybrid Model for Higher Education.(2017,Jun) *International Journal of Engineering and Technology* [en línea]. (9)3. Disponible en https://www.enggjournals.com/ijet/docs/IJET17-09-03-146.pdf
3. N Nghe, P. Janecek y P. Haddawy. "A comparative analysis of techniques for predicting academic performance," *37th Annual Frontiers in Education Conference Global Engineering: Knowledge without Borders, Opportunities without passports*, 2017. DOI 10.1109/FIE.2007.4417993
4. R. Asif, A. Merceron, S. Ali, y N. Haider. "Analyzing undergraduate students performance using educational data mining" *Computers & Education,* vol 113, No 5, pp 177-194, 2017. DOI:10.1016/j.compedu.2017.05.007.
5. A. Shahiri, W. Husain, y R. Abdul. "A review on predicting student's performance using data mining techniques", *Procedia Computer Science,* vol 72, pp 414-422, 2015. DOI:10.1016/j.procs.2015.12.157 , 2015.
6. M Goga, S. Kuyoro, y N. Goga. "A recommender for improving the student academic performance". *Procia -Social and behavioral Sciences*, Vol 180 No 5, pp 1481 – 1488, 2017 DOI:10.1016/j.sbspro.2015.02.296.

7. S. Anil, A. Kurian, s. Dey, S. Saha, y A. Sinha, A. "Genealogy Tree: Understanding Academic Lineage of Authors via Algorithmic and Visual Analysis". *Journal of Scientometric Research*, Vol 7, 120 – 124, 2018. DOI: 10.5530/jscires.7.2.18,

8. C. Beaulac y J. Rosenthal. "Predicting University Students Academic Success and Major using random forests". *Research in Higher Education*, Vol 60, No 7, pp1048 – 1064, 2019.DOI:10.1007/s11162-019-09546-y

9. L. Breiman, J.H. Friedman, R.A. Olshen y C.S. Stone,"Classification and regression trees", 2017. DOI=10.1201%2f9781315139470&partnerID=40&md5=

10. S. Alturki, N. Alturki, y N. Stuckenschmidt, "Using Educational Data Mining To Predict Students Academic Performance For Applying Early Interventions", Journal of Information Technology Education: Innovations in Practice, Vol 20, pp. 121-137, 2021. DOI 10.28945/4835.

11. E.I. Al-Fairouz, M.A. Al-Hagery, "Students performance: From detection of failures and anomaly cases to the solutions-based mining algorithms Allergy", *Asthma and Immunology Research*, Vol 13, No 10, pp. 2895-*2908, 2020.*

12. *R. Montero, Modelos* de regresión lineal múltiple. Documentos de Trabajo en Economía Aplicada. 2016. Disponible en https://www.urg.es/montero/matematicas/regresion_lineal.pdf,.

13. S. Hussain, S. Gaftandzhieva, D. Maniruzzaman, y Z. Muhsin, "Regression analysis of student academic performance using deep learning", *Education and Information technologies*, vol 26 No 01, 2020. DOI:10.1007/s10639-020-10241-0, 2020.

14. P. Strecht, J. Mendes-Moreira, C. Soares, *Merging decision trees: A case study in predicting student performance Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8933, pp. 535-548, 2014. DOI: 10.1007/978-3-319-14717-8_42,.

15. A.B. Urbina-Nájera, A. Téllez-Velázquez y R.C. Barbosa. (2021, Marz), Patterns to Identify Dropout University Students with *Educational Data Mining. Revista Electrónica de Investigación Educativa* [en línea] (23)29. Disponible en https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123517169&doi=10.24320%2fREDIE.2021.23.E29.3918&partnerID=40

16. ICFES. (2019). Instituto Colombiano para la Evaluación de la Educación. Obtenido de https://www.icfes.gov.co/es/web/guest/quien-somos-icfes

17. I. Malá, V. Sládek, y D. Bílková, "Power comparisons of normality tests based on l-moments and classical tests". *Mathematics and Statistics*, Vol 9, No 6, pp. 994-1003, 2021. DOI: 10.13189/ms.2021.090615

18. "R-Studio Team", 2021.[en línea] Disponible en www.rstudio.com

19. Turcios, R.A.S. Student's t (2015,ene-Marz) Uses and abuses. *Revista Mexicana de Cardiología* [en línea] vol 26, No 1, pp. 59-61. Disponible en https://www.scopus.com/inward/record.uri?eid=2-s2.0-84927605287&partnerID=40&md5=aff76bf6cf125f6507040c313868930c

20. [J.G. Lugo-Armenta y L.R., Pino-Fan (2021) Inferential Reasoning Levels for t-Student Statistical. *Bolema - Mathematics Education Bulletin* [en linea] Vol 35 No 71, pp. 1776-1802, DOI: 10.1590/1980-4415V35N71A25.