

**Modelo de Machine Learning para relacionar la formación, tipo de vinculación y cargo de los docentes, con los resultados de las pruebas estandarizadas Saber 11 en los colegios oficiales del departamento Cundinamarca en el año 2023**

**Proyecto de grado**

**Autores:**

José Maximiliano Gómez Torres

*(Jmgomez@unbosque.edu.co)*

Cristina Paola Miranda Escandón

*(Cmirandae@unbosque.edu.co)*

**Director:**

Alejandro Duitama,

*(aduitama@unbosque.edu.co)*

**Maestría en Estadística Aplicada y Ciencia de Datos**

**Departamento de Matemáticas y Estadística**

**Facultad de Ciencias**

**Universidad El Bosque**

Noviembre de 2024

## Tabla de contenido

<b>Resumen</b> .....	<b>4</b>
<b>Abstract</b> .....	<b>5</b>
<b>Justificación</b> .....	<b>6</b>
<b>Introducción</b> .....	<b>7</b>
<b>Revisión de la Literatura</b> .....	<b>8</b>
<b>Metodología</b> .....	<b>11</b>
<b>Conjunto de datos</b> .....	<b>12</b>
<b>Preparación de los datos</b> .....	<b>13</b>
<b>Análisis Exploratorio</b> .....	<b>14</b>
<b>Selección de características</b> .....	<b>16</b>
<b>Normalización</b> .....	<b>16</b>
<b>Balanceo de Datos</b> .....	<b>17</b>
Técnica de Balanceo .....	15
<b>Establecimiento del modelo</b> .....	<b>18</b>
<b>Importancia de las características</b> .....	<b>22</b>
<b>Interpretabilidad del Modelo</b> .....	<b>23</b>
<b>Conclusiones y Discusión</b> .....	<b>24</b>
<b>Bibliografía</b> .....	<b>25</b>

## Índice de gráficos

Gráfica 1. Distribución de los docentes en las ie de cundinamarca.....	14
Gráfica 2. Formación docente por clase .....	13
Gráfica 3-fusión de características .....	13
GRÁFICA 4. CLASIFICACIÓN ICFES DE LOS COLEGIOS DE CUNDINAMARCA .....	17
GRÁFICA 5. CLASIFICACIÓN CON DATOS BALANCEADOS UTILIZANDO ADASYN.....	16
GRÁFICA 6. EVALUACIÓN DE MODELOS DE CLASIFICACIÓN: MÉTRICAS DE RENDIMIENTO PARA DIFERENTES ALGORITMOS USANDO ADASYN .....	18
GRÁFICA 7. MATRIZ DE CONFUSIÓN DE MODELO ET .....	19
GRÁFICA 8. CURVA ROC DE MODELO ET .....	19
GRÁFICA 9. CURVA DE VALIDACIÓN DE MODELO ET .....	21
GRÁFICA 10. IMPORTANCIA DE CARACTERÍSTICAS .....	22
GRÁFICA 11. IMPORTANCIA DE CARACTERÍSTICAS BASADA EN VALORES SHAP POR CLASE" .....	23

## Resumen

Numerosas políticas orientadas al mejoramiento de la calidad en la educación preescolar, básica y media se han centrado en la cualificación de alto nivel de los maestros como una herramienta clave para mejorar el éxito académico de los estudiantes. Sin embargo, la literatura revisada para este trabajo de grado no encuentra evidencia concluyente de una relación significativa entre los títulos de posgrado de los docentes y los resultados de sus estudiantes en pruebas estandarizadas. Investigaciones previas, basadas en estudios que usan modelos de regresión y series de tiempo, han sugerido que una mayor formación de los docentes no necesariamente se traduce en mejores rendimientos académicos de los alumnos.

De forma similar, los enfoques de aprendizaje automático tampoco han identificado la formación avanzada de los docentes como una variable principal que afecte el rendimiento académico de los estudiantes. A pesar de ello, las recientes políticas públicas en Colombia insisten en que la formación de alto nivel de los docentes es un componente determinante en la mejora de los aprendizajes, particularmente en los niveles de preescolar, básica y media.

En este contexto, el presente trabajo busca aportar sobre esta hipótesis de la política educativa: ¿realmente el nivel de formación, el tipo de vinculación y en rol de la institución educativa -IE de los docentes, influye de manera significativa en los resultados académicos, en los términos de las pruebas estandarizadas?

Este trabajo propone un modelo basado en algoritmos de aprendizaje automático que predice la clasificación ICFES de los colegios oficiales en Cundinamarca. Para ello, se construyó una base de datos que integra características clave del perfil docente, incluyendo su nivel de formación (desde normalistas superiores hasta doctorados), su tipo de vinculación laboral al magisterio y su rol en la institución educativa. Estos atributos fueron analizados junto con la clasificación obtenida por cada colegio en las pruebas ICFES.

Los resultados revelan que el modelo predictivo logró un rendimiento moderado, donde la formación de alto nivel de los docentes en niveles de especialización y maestría es una de las características más influyentes en la predicción de los resultados de las IE. Nuestro trabajo sugiere además que otros factores, como ser docente en propiedad y docente de aula, pueden tener un impacto considerable en la clasificación de los planteles educativos.

Esperamos que este trabajo de grado represente una contribución al campo, al aplicar técnicas

de aprendizaje automático para predecir los resultados de la clasificación ICFES de los colegios oficiales, usando características de formación, vinculación y rol de los docentes en las IE. Los hallazgos, que deben corroborarse con la información de otras entidades territoriales, podrían guiar futuras investigaciones y aportes sobre las verdaderas palancas para mejorar la calidad educativa en Colombia.

### **Contribuciones:**

1. Se ha reunido un conjunto de datos enfocado en predecir el desempeño de las IE del departamento, a partir de la formación avanzada, el tipo de vinculación y el rol de los docentes.
2. Este trabajo analiza las características de los docentes que influyen en la calidad de las IE, medida a través de pruebas estandarizadas, dejando de lado factores específicos del rendimiento individual de los estudiantes y su entorno.
3. Se ofrecen elementos para futuros modelos que profundicen en los factores que afectan la calidad educativa de las IE, usando las características docentes como variable central. De esta manera, se contribuye al entendimiento de las políticas públicas actuales y futuras, orientadas a fortalecer la calidad en los niveles de educación preescolar, básica y media.

### **Abstract**

Numerous policies aimed at improving quality in preschool, elementary, and secondary education have focused on high-level teacher qualification as a key tool for enhancing students' academic success. However, the literature reviewed for this thesis finds no conclusive evidence of a significant relationship between teachers' postgraduate degrees and their students' results on standardized tests. Previous studies, using regression models and time series, have suggested that increased teacher training does not necessarily translate into higher academic performance among students.

Similarly, machine learning approaches have not identified advanced teacher training as a primary variable affecting students' academic outcomes. Nevertheless, recent public policies in Colombia insist that high-level teacher training is a crucial component for improving student learning, especially at the preschool, elementary, and secondary levels.

In this context, the present study seeks to explore this educational policy hypothesis: Does the level of teacher training, employment type, and role within the educational institution (IE) significantly influence academic results, as measured by standardized tests?

This work proposes a model based on machine learning algorithms to predict the ICFES classification of public schools in Cundinamarca. To achieve this, a database was built that integrates key characteristics of the teacher profile, including their level of training (from normalista superior to doctoral levels), their type of employment within the teaching workforce, and their role in the educational institution. These attributes were analyzed alongside the classification each school received on the ICFES tests.

The results reveal that the predictive model achieved a moderate performance, where teachers' advanced training at the specialization and master's levels is one of the most influential characteristics in predicting the outcomes of IE. Our work also suggests that other factors, such as being a tenured teacher and classroom teacher, can have a substantial impact on school classifications.

We hope that this thesis contributes to the field by applying machine learning techniques to predict ICFES test results, using characteristics of teacher training, employment type, and roles within the IE. The findings, which should be corroborated with data from other regions, could guide future research and insights on the true levers for improving educational quality in Colombia.

## **Justificación**

El 29 de septiembre de 2023, el presidente Gustavo Petro dijo en la localidad bogotana de Engativá: "Tiene que haber un Gobierno nacional que sea amigo de la tesis de mejorar la educación y producir un gran salto hacia adelante. La prioridad es la educación, la primera partida del presupuesto nacional. Desde cuándo fui alcalde de la ciudad iniciamos un programa de formación en las mejores universidades para cientos de maestros y maestras; ahora debemos extender este programa a nivel nacional y vamos a educar a 25.000 maestros con cursos de postgrado para que puedan ayudar a elevar la calidad de la educación en Colombia" (MEN, 2023). Esto se empezó a materializar en la convocatoria de Poder Pedagógico Popular, realizada en febrero de 2024 por el Ministerio de Educación Nacional y el ICETEX, para cursar programas de maestría y doctorado.

Si bien, los anuncios de formación de alto nivel de los docentes de colegios oficiales se repiten cada 4 años a nivel nacional y territorial, donde los gobiernos destinan presupuestos nada despreciables para alcanzar las metas de formación docente que se trazan, recurriendo a la premisa de que entre más docentes con maestrías y doctorados habrá mejores resultados en la calidad de la formación de los estudiantes de básica y media, no parece existir en el país mediciones que respalden este supuesto. Mas importante aún, a pesar de la mejora en la cualificación del magisterio que se viene extendiendo a escala nacional, dignificando a los docentes, mejorando su escala salarial, los resultados en pruebas estandarizadas y en mediciones internacionales, no parecen mejorar en los últimos años.

En este contexto, este trabajo busca aportar evidencia, evaluando si la cualificación, el tipo de vinculación y el rol de los docentes, efectivamente contribuye a elevar la calidad educativa en Colombia. Nuestros resultados aportan información valiosa para la toma de decisiones, orientando la política educativa hacia enfoques basados en evidencia que optimicen el uso de recursos y dirijan los esfuerzos hacia aquellos factores que realmente impactan el rendimiento académico.

## **Introducción**

La calidad de la educación depende de muchas causas, como las condiciones socioeconómicas de los estudiantes, estrategias de enseñanza y aprendizaje, la capacitación de los maestros, los métodos de evaluación utilizados, la coherencia entre los diferentes niveles educativos, la infraestructura y capacidad de las instituciones, así como el marco conceptual y legal que sustenta el sistema educativo (Contreras, Tarazona, & y otros, 2023).

Diversos estudios han hecho aportes significativos para avanzar en la comprensión de los factores que influyen en los aprendizajes de los estudiantes y la analítica de datos contribuye cada vez más a dar luces para que los tomadores de decisiones mejoren la formulación de sus planes de acción. (Long & Siemens, 2011). Sin embargo, parecen existir más estudios centrados en evaluar las causas individuales del éxito académico de los estudiantes que aquellos centrados en comprender el desempeño institucional y de sus factores determinantes, entre ellos el número y características de los docentes. (Hammed, Mahboob, & Mushtaq, 2022)

Cuando se trata de impactar los aprendizajes de los estudiantes, los documentos de política pública suelen tomar en consideración el desarrollo profesional de los docentes, desde su formación inicial, la formación a lo largo de su quehacer en el aula y la formación de alto nivel,

es decir especializaciones, maestrías y doctorados. (MEN, 2013). La política de calidad en educación preescolar básica y media de nuestro país, ha venido creciendo en programas que apuestan cada vez más a brindar becas, que cualifiquen a los docentes en formación de alto nivel (MEN, 2008); (MEN, 2011); (MEN, 2023). Por su parte, la Secretaría de Educación de Cundinamarca consideró fundamental formar a los docentes del departamento usando la información provista por Saber 11, para identificar las necesidades específicas en cada IE pues se reconoció que estos resultados ofrecen una perspectiva para orientar la formación docente necesaria y detectar oportunidades de mejora en los programas de formación (SEC, 2021)

La revisión de la literatura examina cómo, los estudios cuantitativos no han encontrado una relación directa entre la formación de alto nivel de los docentes y los resultados académicos de los estudiantes. Estudios en contextos internacionales, como los de (Buddin & Zamarro, 2009) y (Clotfelter, Ladd, & Vigdor, 2007), señalan que las credenciales avanzadas de los docentes, como maestrías o doctorados, no tienen un impacto significativo en el rendimiento estudiantil. Además, revisiones sistemáticas recientes sobre el uso del aprendizaje automático para predecir el desempeño escolar, como las de (Issah & y otros, 2023) y (Contreras, Tarazona, & y otros, 2023), detectan que los modelos otorgan mayor relevancia a características sociodemográficas, de entorno institucional y de desempeño previo de los estudiantes en comparación con la cualificación docente en la predicción del éxito estudiantil. (Miller, Murnane, & Willett, 2008).

Este trabajo propone un modelo de aprendizaje automático que predice la clasificación ICFES de los colegios de Cundinamarca utilizando como variables la formación, el tipo de vinculación y cargo de los docentes. Se describen los métodos de recolección y análisis de datos, las técnicas de modelado empleadas, y se discuten los resultados obtenidos.

Finalmente, se presentan las conclusiones y recomendaciones que pueden servir como base para futuros estudios, destacando las implicaciones prácticas de los hallazgos y aportando un análisis crítico sobre el impacto real de la cualificación docente en el éxito académico de los estudiantes.

## **Revisión de la Literatura**

Desde el Plan Sectorial de Educación 2006-2010, Colombia ha planteado el desarrollo profesional de los docentes como una estrategia fundamental para mejorar la calidad de la educación preescolar, básica y media (MEN, 2013). Según el Ministerio de Educación

Nacional (MEN, 2008), "la calidad se materializa en las instituciones educativas, donde un equipo de docentes, bajo el liderazgo del rector o director y su equipo directivo, realiza las actividades pedagógicas establecidas en el proyecto educativo institucional para que los estudiantes desarrollen sus competencias. De ahí la importancia de contar con un cuerpo docente altamente calificado". Para lograrlo, el Plan Decenal consideró necesario articular los niveles de formación inicial, pregrado, posgrado y la formación permanente de los maestros, destacando la importancia de la formación continua como pilar esencial para su crecimiento profesional.

No obstante, desde la perspectiva del sistema de formación docente en Colombia, no se aborda claramente la relación directa entre la formación de alto nivel y los resultados académicos de los estudiantes. Más bien, los títulos de especialización, maestría y doctorado proyectan a los docentes hacia espacios de investigación, innovación y gestión educativa, para crear soluciones ante los desafíos del sistema educativo y el sistema de formación de educadores (MEN, 2013). Según el Ministerio, los programas de formación avanzada tienen un enfoque más amplio, buscando que los docentes no solo sean mejores en el aula, sino que también participen activamente en la generación de conocimiento, el diseño de políticas educativas y la implementación de estrategias pedagógicas innovadoras.

En Bogotá, bajo el marco del plan de desarrollo "Bogotá Humana" entre 2010 y 2014, se implementó el proyecto "Construcción de Saberes: Educación Incluyente, Diversa y de Calidad para Disfrutar y Aprender", que estableció como meta que el 30% de los docentes y directivos docentes accedieran a formación de excelencia en doctorados, maestrías y especializaciones (Alcaldía Mayor de Bogotá, 2012). Este fue uno de los proyectos más ambiciosos a nivel nacional en términos de formación avanzada para los docentes de colegios oficiales, ya que buscaba mejorar la calidad educativa en la capital al fortalecer la capacidad pedagógica y de gestión de sus educadores.

Parece existir consenso en los programas de gobierno de la última década en que los docentes bien formados están mejor equipados para planificar y ejecutar estrategias de enseñanza efectivas. Una mayor formación permite a los maestros comprender mejor las necesidades de los estudiantes, adaptar su enfoque pedagógico y utilizar métodos que promuevan un aprendizaje significativo. Estas habilidades contribuyen directamente a la calidad del proceso educativo e inspiran a los estudiantes, fomentando su interés y participación en el proceso educativo. La calidad de la enseñanza influye directamente en la motivación de los estudiantes

y en su disposición para aprender (SED, 2018). Esto subraya la importancia de un enfoque integral en la formación docente, donde no solo se considera la obtención de títulos, sino también el desarrollo de competencias pedagógicas clave para un aprendizaje inclusivo y efectivo. (SED, 2018)

Sin embargo, a pesar de los enunciados de la política, diversos estudios que incorporan técnicas estadísticas y de aprendizaje automático para orientar la toma de decisiones en el ámbito educativo no han identificado la cualificación docente ni su rol en la escuela como variables dominantes para explicar los resultados académicos de los estudiantes o las clasificaciones de los planteles educativos. Para (Buddin & Zamarro, 2009), la calidad de los maestros es un factor clave en el éxito académico de los estudiantes, pero pocas características específicas de los maestros influyen directamente en los resultados en el aula. En su estudio sobre un distrito escolar de Los Ángeles, examinaron si los puntajes en pruebas de licencia de maestros, junto con otros atributos como la experiencia y los antecedentes educativos, se relacionaban con el rendimiento de los estudiantes de primaria. Utilizaron un panel de datos que incluyó variables como el género, etnia, educación de los padres, características del aula, pruebas de maestros, educación y experiencia. Una de sus conclusiones fue que los estudiantes con maestros con maestría o doctorado no obtuvieron mejores ni peores resultados en lectura o matemáticas en comparación con aquellos con maestros sin títulos avanzados.

Otros autores llegan a conclusiones similares. (Clotfelter, Ladd, & Vigdor, 2007), utilizando datos longitudinales de Carolina del Norte, encontraron que, aunque los maestros reciben aumentos salariales por obtener títulos de posgrado, estos no tienen un efecto significativo en el rendimiento de los estudiantes. En algunos casos, incluso, el efecto fue negativo. Esto plantea preguntas sobre la efectividad de recompensar los títulos avanzados sin un análisis profundo de cómo impactan realmente en el aula. (Staiger & Rockoff, 2010), así como (Croninger, Rey Arroz, Rathbun, & Nishio) y (Rivkin & Schiman, 2015), también concluyeron que la cualificación de alto nivel no es un predictor robusto del éxito académico.

En una revisión sistemática de 84 estudios sobre el uso del aprendizaje automático para identificar los factores que influyen en el rendimiento académico, (Issah & y otros, 2023) concluyeron que los factores más frecuentemente utilizados en la predicción del desempeño estudiantil son el rendimiento académico previo (34.2%), los datos demográficos (22.8%), los comportamientos (17.5%), los aspectos psicológicos (8.8%), los antecedentes familiares (8.8%) y las características del entorno escolar (7.9%). Ninguno de estos estudios identificó la

formación docente como un factor relevante, lo que resalta una desconexión entre las políticas de formación docente y los resultados académicos.

Por otro lado, (Contreras, Tarazona, & y otros, 2023) también identificaron categorías similares en su estudio sobre el rendimiento académico en educación superior. Revisaron múltiples trabajos entre 2010 y 2020 sobre el uso de machine learning y deep learning en la predicción del rendimiento académico. En su análisis, destacaron factores académicos como los resultados de pruebas estandarizadas, los promedios académicos y los resultados de exámenes de admisión, así como factores sociodemográficos como la edad, el género, los ingresos familiares y el nivel educativo de los padres. Asimismo, identificaron factores psicológicos, que incluyen la motivación, la dedicación al estudio y las competencias cognitivas, así como factores de gestión académica y el entorno escolar, que también juegan un papel crucial en el rendimiento estudiantil.

Finalmente, expertos en educación han reflexionado sobre la eficacia de las maestrías en educación en Colombia, destacando que, a pesar del aumento significativo en la formación docente con estudios de posgrado, la calidad de la educación básica no ha mejorado, pues la política pública no se enfoca en factores importantes como el currículo, el liderazgo pedagógico y la participación comunitaria. A pesar de las grandes inversiones del país en formación docente, los resultados en pruebas internacionales y nacionales, como PISA y SABER 11, muestran que las brechas entre colegios públicos y privados siguen aumentando.

(De Zubiría, 2020) sugiere que promover la formación en grupos colectivos de docentes dentro de los colegios, desarrollar maestrías contextualizadas e in situ que aborden los problemas específicos de cada institución y mejorar la formación inicial de los docentes enfocándose en competencias esenciales como la lectura crítica y el razonamiento numérico son indispensables para lograr una transformación real en el sistema educativo colombiano.

## **Metodología**

La metodología seguida en este trabajo consistió en la recopilación, preparación y procesamiento de datos, modelado y evaluación de resultados. Para la construcción del conjunto de datos, se utilizaron bases oficiales de la Secretaría de Educación de Cundinamarca y el ICFES correspondientes al año 2023.

En la etapa de preparación de datos, se realizaron procesos de limpieza, así como de transformación y normalización de datos para asegurar la consistencia y homogeneidad entre las variables. Posteriormente, se aplicaron técnicas de balanceo de datos (como SMOTE, ADASYN y sub y sobre muestreo) para corregir la desproporción entre las clases y evitar el sesgo hacia las clases mayoritarias. Además, se utilizaron métodos de selección de características para identificar las variables predictoras más relevantes.

Para el modelado, se entrenaron varios clasificadores de aprendizaje automático, entre ellos árboles de decisión, bosques aleatorios y máquinas de vectores de soporte. La evaluación del rendimiento de cada modelo se realizó mediante métricas como precisión, sensibilidad, entre otras.

Todas las etapas de análisis, preparación y modelado de datos se implementaron en Google Colab 1.0.0 y Python 3.10.12 para la manipulación, modelado y visualización de datos, respectivamente.

### **Conjunto de datos**

Para este trabajo, se emplearon dos bases de datos oficiales y anonimizadas. La primera proviene de la Secretaría de Educación de Cundinamarca y contiene información detallada sobre la formación y tipo de vinculación de los docentes de las IE oficiales del departamento. Esta base incluye datos sobre el nivel de formación, que varía desde bachiller y técnico profesional hasta niveles avanzados como magíster y doctorado, en diferentes áreas de estudio.

También se registran aspectos del tipo de vinculación (en encargo, provisional, periodo de prueba, propiedad, planta temporal), la función en el colegio (coordinador, docente de aula, orientador, tutor, rector, entre otros), el nivel educativo asignado (preescolar, primaria, secundaria o media), así como el sexo, edad y la institución educativa correspondiente.

La segunda base de datos proviene del ICFES y ofrece información sobre la clasificación Saber 11 de cada IE, así como el número de estudiantes que presentaron la prueba en 2023. Esta clasificación se calcula en base a los resultados obtenidos en las pruebas Saber 11 de los últimos tres años, que evalúan a los estudiantes en cinco áreas clave: lectura crítica, matemáticas, estudios sociales, ciencias naturales e inglés. Según el (ICFES, 2017), la clasificación se realiza considerando únicamente el 80% superior de los estudiantes con los puntajes más altos,

excluyendo a los repitentes y a los estudiantes con discapacidad, para evitar sesgos. Las categorías de clasificación van desde A+ (con un índice global entre 0.77 y 1), A (0.72 a 0.77), B (0.67 a 0.72), C (0.62 a 0.67) y D (menos de 0.62). Esta base de datos también proporciona el número total de estudiantes que presentaron la prueba en 2023, lo que permite contextualizar los resultados académicos.

Ambas bases de datos se integraron mediante el uso del código DANE, un identificador único e intransferible que permite identificar a cada sede educativa de manera inequívoca en Colombia. Este proceso de integración resultó en una base de datos consolidada que incluye 273 instituciones educativas de Cundinamarca, combinando las características de formación y tipo de vinculación docente con la clasificación Saber 11 y el número de estudiantes evaluados. Esta base de datos consolidada permite una exploración más detallada de cómo las características institucionales y las condiciones del personal docente impactan en los logros académicos de los estudiantes en la región de Cundinamarca.

### **Preparación de los datos**

Durante el proceso de limpieza de datos, identificamos que cuatro instituciones educativas no contaban con clasificación ICFES. La IED Barroblanco de Bojacá y la IED Departamental Gazaduje de Medina no cuentan con media, por tanto, sus estudiantes deben transitar a otra IE para finalizar su formación. Adicionalmente, aunque la IE Nacionalizado de Fúquene y el Instituto Nacional de Promoción Social del municipio de Villeta ofrecen educación media, el ICFES no las clasificó en 2023. Se tomó la decisión de eliminarlas de la base de datos.

Posteriormente, se tomó la decisión metodológica de filtrar la base de datos para incluir únicamente la información de los docentes asignados a básica secundaria y media. Esta decisión se fundamenta en la alta movilidad docente en el departamento, lo que hace poco probable que los docentes que actualmente trabajan en básica primaria y poseen formación avanzada hayan influido de manera directa en el proceso formativo de los estudiantes que presentaron la prueba Saber 11 en 2023. Aunque reconocemos que esta selección no resuelve por completo la problemática de movilidad y temporalidad, es una suposición menos exigente, al considerar que los estudiantes podrían haber estado bajo la influencia de los docentes durante los últimos seis años, en lugar de once.

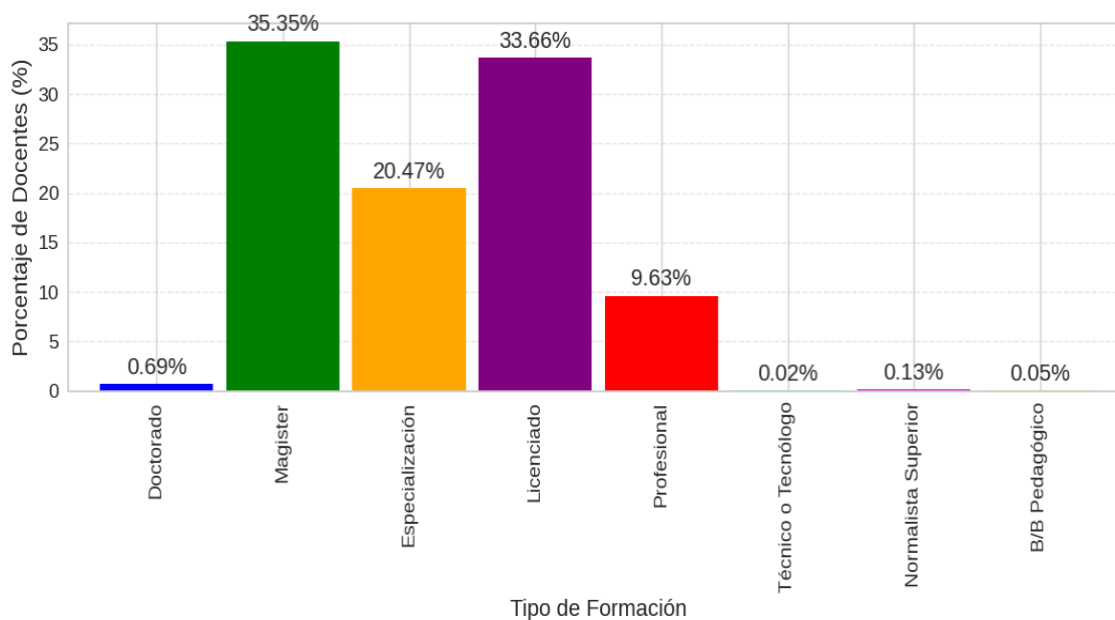
Es importante mencionar que al seleccionar solo a los docentes de básica secundaria y media, se reduce la complejidad del análisis al acotar la influencia docente a los años más cercanos al periodo de las pruebas Saber 11. De esta forma, se minimiza la incertidumbre sobre la permanencia de los docentes en una misma IE y se asegura un enfoque más coherente sobre la influencia de la formación académica de los docentes en los resultados de los estudiantes.

En términos prácticos, la base de datos inicial contenía información sobre 10,389 docentes. Tras aplicar el filtro metodológico, la base final utilizada en el análisis incluyó 5,341 docentes y directivos docentes, todos ellos asignados a los grados de educación secundaria y media. Este enfoque permite un análisis más preciso de la relación entre la formación, la vinculación y el cargo de los docentes y el rendimiento académico de los estudiantes en Cundinamarca.

### Análisis Exploratorio

Brevemente presentamos la composición de la formación docente en las IE de Cundinamarca, así como de la vinculación y el cargo.

**Gráfica 1. Distribución de los docentes en las IE de Cundinamarca según formación**

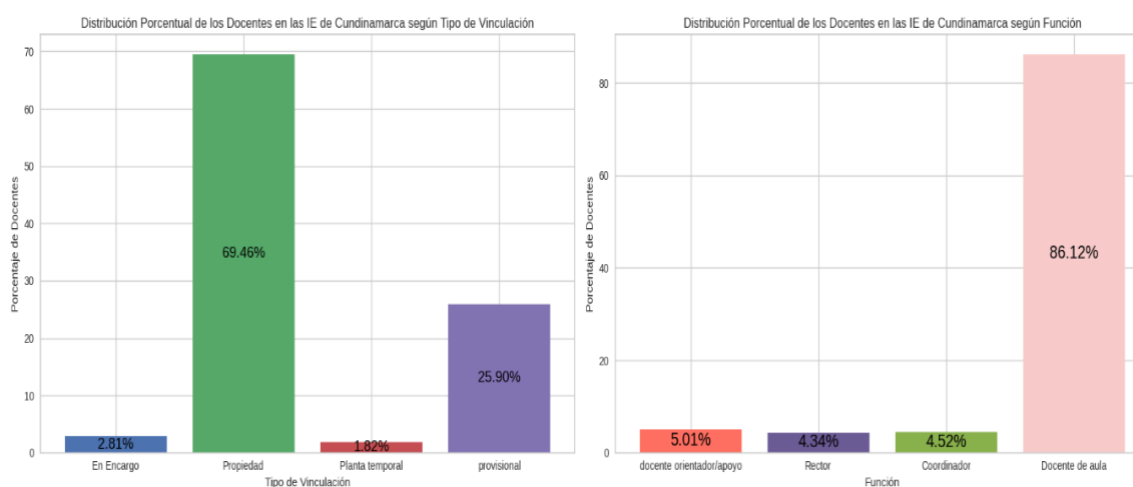


La mayoría de los docentes son Magister (35.35%) y en la segunda formación más común es la de Licenciado (33.66%). Hay que decir que en la base de datos solo se registra la última formación del docente, así que todos los docentes con formación avanzada necesariamente son también licenciados o profesionales universitarios. La Especialización es el siguiente tipo

de formación avanzada más común con un 20.47%. En contraste, la formación Profesional representa un 9.63%, mientras que los tipos de formación como Doctorado (0.69%), Normalista Superior (0.13%), Técnico o Tecnólogo (0.02%), y Bachiller / Bachiller Pedagógico (0.05%) tienen una representación significativamente menor.

Por otra parte, la gráfica 2 izquierda, muestra la distribución porcentual de los docentes en función de su tipo de vinculación. La categoría Propiedad es la más prevalente, representando el 69.46% del total, lo cual indica que la mayoría de los docentes tienen una vinculación estable y permanente en el sistema educativo del departamento. Le sigue la categoría Provisional con un 25.90%, que agrupa a los docentes que ocupan su posición de manera temporal o provisional. Las categorías En Encargo y Planta Temporal son minoritarias, con porcentajes de 2.81% y 1.82%, respectivamente. Esta distribución sugiere una estructura laboral predominantemente estable en el sector educativo de Cundinamarca, aunque un cuarto de los docentes tiene una vinculación provisional.

### Gráfica 2. Distribución de los docentes en las IE de Cundinamarca según tipo de vinculación y función



En la parte derecha de la gráfica, se presenta la distribución de los docentes según su tipo de función en las instituciones educativas. La mayoría de los docentes, un 86.12%, son Docentes de aula, lo que subraya su rol central en la enseñanza directa a los estudiantes. Las funciones de "Coordinador", "Docente orientador/apoyo" y "Rector" representan el 4.52%, 5.01% y 4.34%, respectivamente, indicando que solo una pequeña proporción de los docentes ocupa roles de apoyo, liderazgo o administración dentro de las instituciones. Esta distribución pone de manifiesto la preponderancia de los docentes en funciones de enseñanza directa en comparación con aquellos en roles de apoyo o liderazgo.

### Selección de características

Se agruparon características con muy baja frecuencia en la base de datos y que conceptualmente pueden considerarse similares. A continuación, sintetizamos las fusiones realizadas.

### Gráfica 2-Fusión de Características

Nuevas Características	Características Originales Fusionadas
Planta temporal	Planta temporal NEE, Planta Temporal PTA
Propiedad	Período de prueba, Propiedad
Provisional	Provisional vacante definitiva, provisional vacante temporal
Docente orientador/apoyo	Docente con funciones de apoyo, Docente Orientador, Docente tutor
Rector	Rector Escuela Normal Superior, Rector Institución Educativa Básica completa, Rector Institución Educativa completa
Bachiller	Bachiller, Bachiller Pedagógico
Doctorado	Doctorado en Educación, Doctorado en otra áreas
Especialización	Especialización en Educación, Especialización en otras áreas, Posgrado en Educación
Licenciado	Licenciado, No definido, Maestro
Magister	Magister en Educación, Magíster en otras áreas
TyT	Técnico o Intermedia Profesional, Tecnólogo

### Normalización

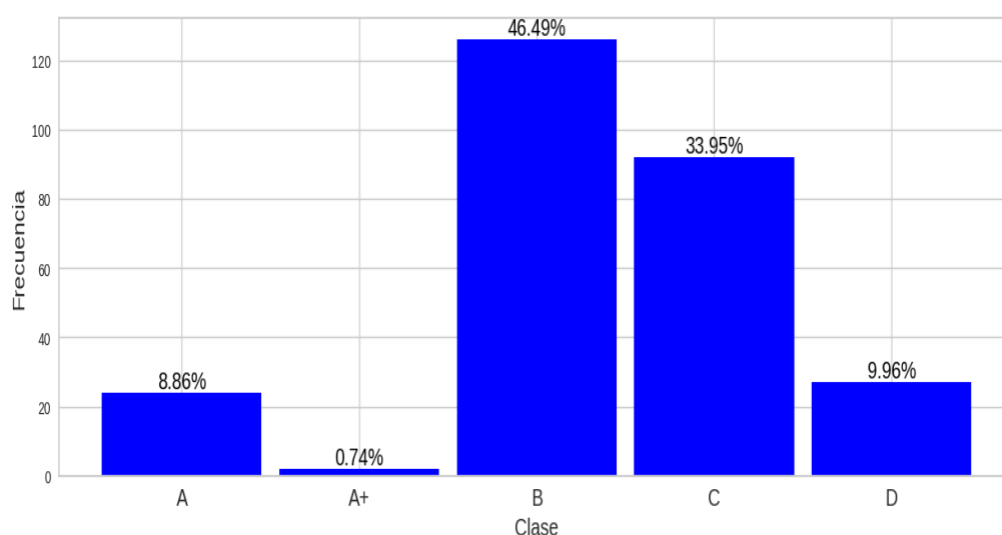
Los valores de la base de datos están significativamente influenciados por el tamaño de las Instituciones Educativas Departamentales (IED). Los colegios con un mayor número de estudiantes tienden a contar con más personal docente, lo que puede influir en los análisis y resultados posteriores. Un claro ejemplo de esta disparidad es que 76 colegios concentran el 50% del total de docentes, mientras que los otros 198 abarcan el otro 50%. Instituciones como la IE Miguel Samper (Guaduas), con 142 docentes, la IE Departamental Serrezuela (Madrid), con 135, o la IE María Teresa Ortiz (Madrid), con 131, contrastan con otras que apenas cuentan con 10 docentes, como la IE El Hato (Carmen de Carupa), Camancha y Cuibuco (San Cayetano), o la IE Integrado de Santa Rosa (Ubalá).

Esta variación en la cantidad de docentes introduce una heterogeneidad que requiere un tratamiento cuidadoso. En este trabajo, hemos aplicado la normalización MinMax para ajustar las diferencias y evitar que el tamaño de las IEDs sesgue los resultados, facilitando así una interpretación más equitativa y precisa de los datos.

### Balanceo de Datos

La base de datos tiene un fuerte desbalance. Los colegios Clase B (46.49%), es la clase mayoritaria representando casi la mitad del total. La Clase C (33.95%), también es bastante común, con un tercio del total. Las Clases D (9.96%) y A (8.86%): Son minoritarias pero tienen una presencia significativa y la Clase A+ (0.74%) es extremadamente minoritaria, lo que impide que el modelo aprenda patrones significativos sobre esta clase y que nos llevó a proponer estrategias de balanceo SMOTE, ADASYN, decantándonos por un submuestreo de las clases mayoritarias y sobre muestreo de las minoritarias a 90, debido a los mejores resultados en la modelación

**Gráfica 3. Clasificación Icfes de los Colegios de Cundinamarca**



En este punto queremos mencionar que se valoró la posibilidad de agrupar la clase A+ y A, sin embargo, no encontramos similitudes evidentes entre las características de los 2 planteles educativos que nos facilitarían tomar esa decisión.

Adicionalmente, clasificar a las Instituciones Educativas no tiene solamente como objetivo integrarlas en un sistema de evaluación estandarizada que permita evaluar las competencias de los estudiantes de educación media sino también proporcionar información sobre la calidad de la educación en el país (ICFES, 2023) y atendiendo al objetivo del trabajo, consideramos mantener la clase A+ de forma independiente y relevarla.

## Establecimiento del modelo

La siguiente modelación se realizó usando Pycaret versión 3.3.2, la cual es una biblioteca de Python de código abierto que simplifica el ciclo de vida de la ciencia de datos y la ingeniería de aprendizaje automático. Es una alternativa ligera y fácil de usar, permitiendo a los usuarios prototipar rápidamente modelos de ML con pocas líneas de código. (Pycaret.org, 2020)

## Gráfica 5. Evaluación de Modelos de Clasificación: Métricas de Rendimiento para Diferentes Algoritmos

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
<b>et</b>	Extra Trees Classifier	0.7368	0.9470	0.7368	0.7299	0.7150	0.6709	0.6797	0.3690
<b>rf</b>	Random Forest Classifier	0.7209	0.9401	0.7209	0.6904	0.6874	0.6511	0.6658	0.3130
<b>xgboost</b>	Extreme Gradient Boosting	0.7094	0.9041	0.7094	0.7019	0.6904	0.6367	0.6447	0.1410
<b>lightgbm</b>	Light Gradient Boosting Machine	0.7094	0.9066	0.7094	0.6744	0.6768	0.6367	0.6474	0.6210
<b>gbc</b>	Gradient Boosting Classifier	0.7054	0.0000	0.7054	0.6858	0.6791	0.6317	0.6426	0.6080
<b>dt</b>	Decision Tree Classifier	0.6774	0.7986	0.6774	0.6655	0.6476	0.5967	0.6068	0.0730
<b>knn</b>	K Neighbors Classifier	0.5497	0.8155	0.5497	0.5118	0.5122	0.4370	0.4510	0.0540
<b>nb</b>	Naive Bayes	0.5297	0.7979	0.5297	0.4351	0.4490	0.4116	0.4410	0.0650
<b>lda</b>	Linear Discriminant Analysis	0.5297	0.0000	0.5297	0.4989	0.4922	0.4119	0.4236	0.0670
<b>ridge</b>	Ridge Classifier	0.5057	0.0000	0.5057	0.4475	0.4451	0.3817	0.3995	0.0820
<b>lr</b>	Logistic Regression	0.5017	0.0000	0.5017	0.4460	0.4496	0.3767	0.3915	0.8420
<b>svm</b>	SVM - Linear Kernel	0.4940	0.0000	0.4940	0.4314	0.4293	0.3672	0.4030	0.0860
<b>ada</b>	Ada Boost Classifier	0.4818	0.0000	0.4818	0.4293	0.4119	0.3524	0.3819	0.1470
<b>qda</b>	Quadratic Discriminant Analysis	0.2031	0.0000	0.2031	0.0413	0.0687	0.0000	0.0000	0.0440
<b>dummy</b>	Dummy Classifier	0.1992	0.5000	0.1992	0.0397	0.0662	0.0000	0.0000	0.0390

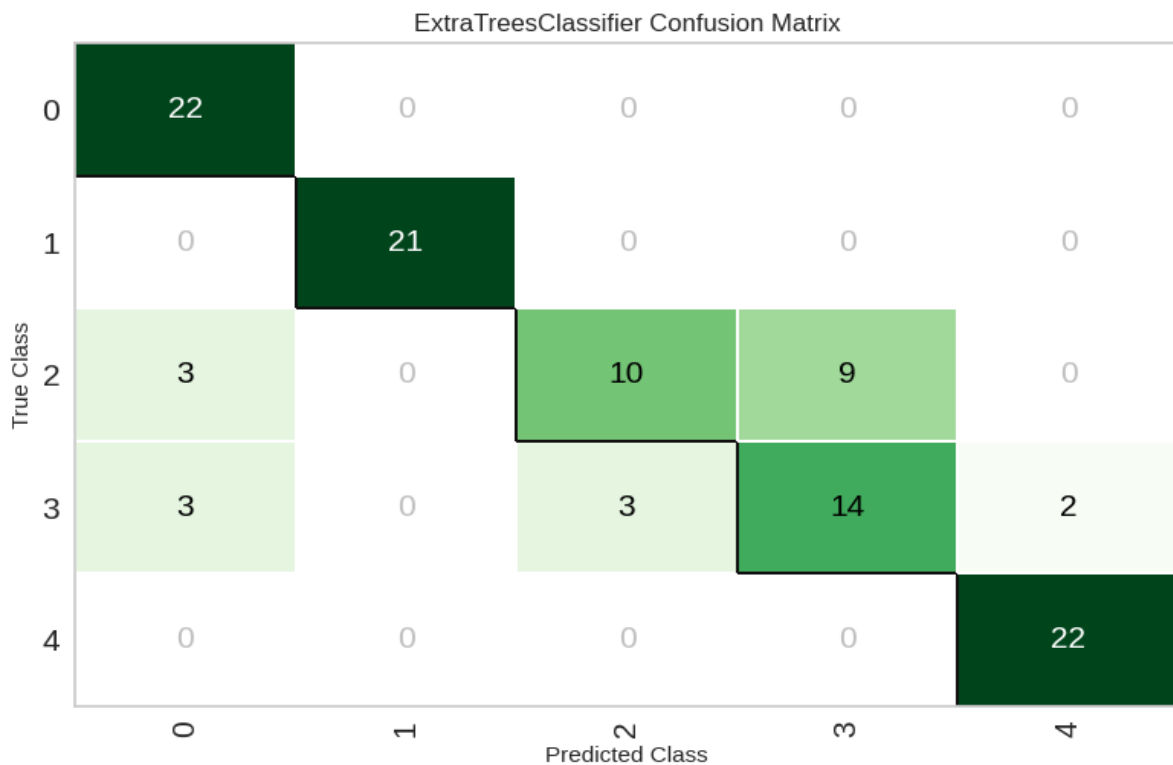
Los mejores modelos están basados en árboles de decisión aleatorios. El modelo Extra Trees Classifier (ET) muestra el mejor rendimiento en comparación con los otros modelos de la tabla, así que se seleccionó y se procedió a realizar el afinamiento de parámetros.

El resultado del afinamiento fue un rendimiento medio con una Accuracy de 0.6513, lo que significa que el modelo clasifica correctamente aproximadamente el 65% de las instancias. El AUC de 0.8943 es bastante alto, indicando que el modelo es eficaz en la discriminación entre las clases, especialmente en el contexto de un modelo de clasificación multiclase. El Recall y la Precision, ambos en 0.6513 y 0.5886 respectivamente, muestran que el modelo es relativamente equilibrado en la detección de verdaderos positivos y la precisión de sus predicciones, aunque sin duda tiene que ser mejorado. En conjunto, estos resultados reflejan

que el modelo tuneado tiene un rendimiento moderado, pero es necesario buscar mayores consistencias en la clasificación y precisión en diversos conjuntos de datos.

A continuación, detallaremos algunos resultados.

**Gráfica 4. Matriz de confusión de modelo ET**

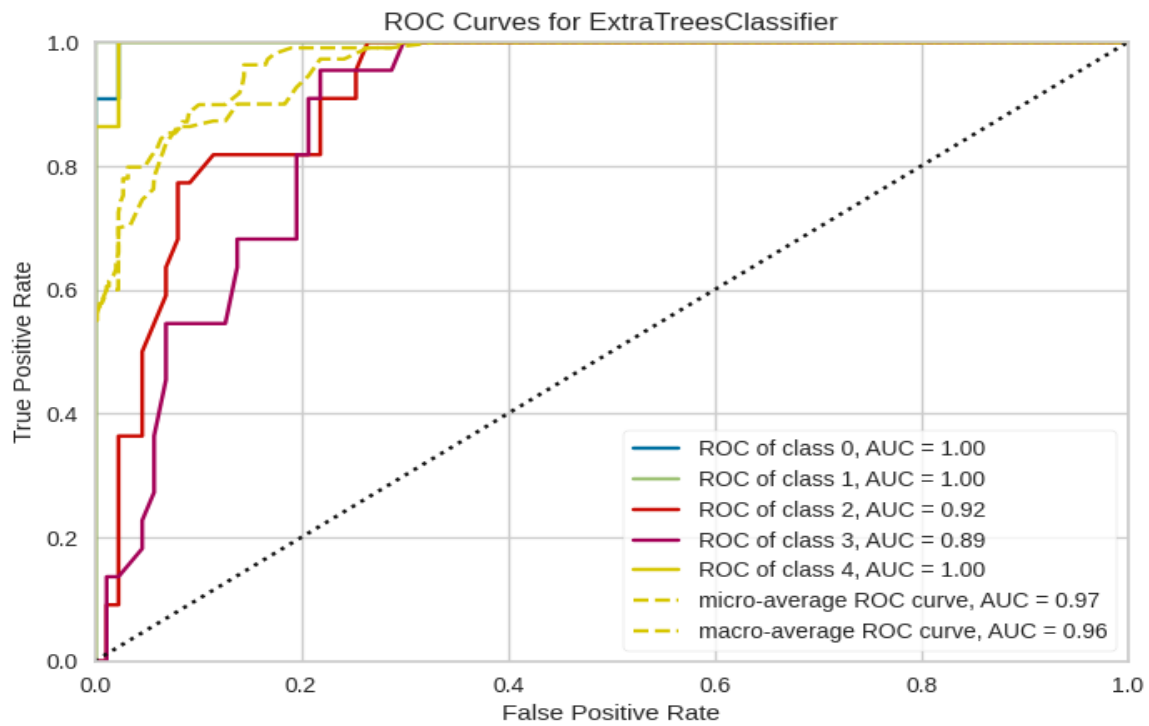


El modelo muestra un rendimiento general moderado, ya que la mayoría de las predicciones se concentran en la diagonal principal, lo que indica una buena capacidad de clasificación para las cinco clases: A, A+, B, C y D. Sin embargo, se observan ciertos errores de clasificación en las clases B y C, que son confundidas con mayor frecuencia con otras clases. Por ejemplo, la clase B tiene 3 instancias incorrectamente clasificadas como clase C, y la clase C presenta 9 instancias clasificadas como clase B. Esta tendencia indica que el modelo encuentra dificultades para diferenciar entre las clases intermedias, probablemente debido a la similitud en sus características. A pesar de estas confusiones, el modelo mantiene una precisión alta en las clases A, A+ y D, lo que refuerza su eficacia en estos segmentos del conjunto de datos.

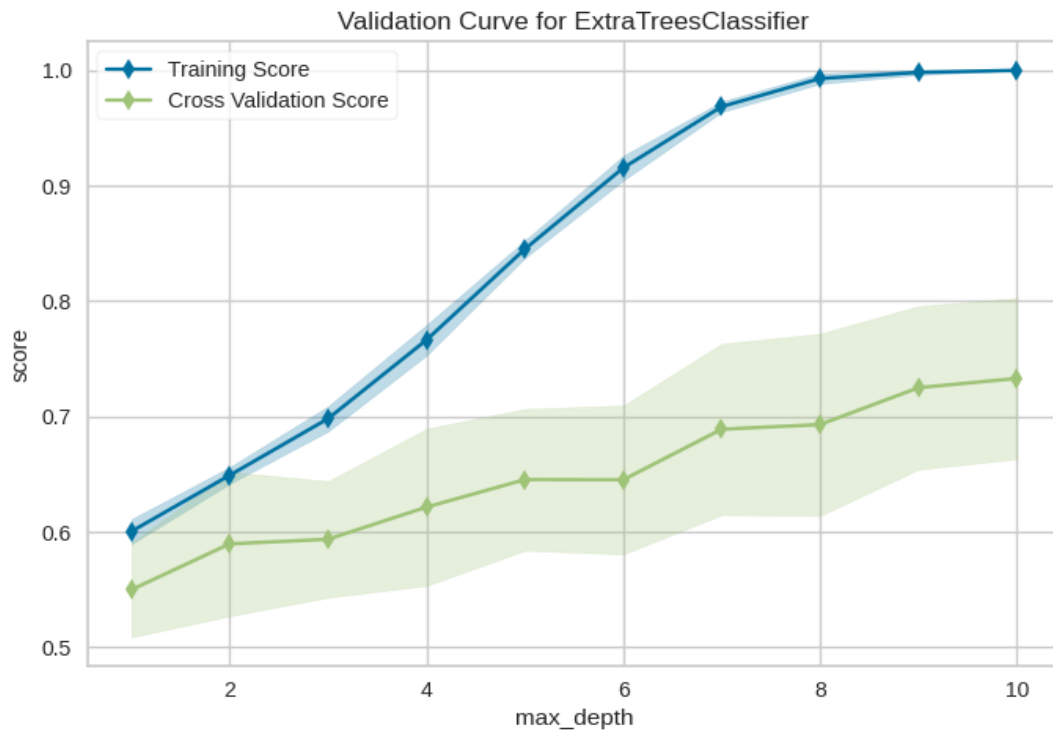
Las curvas ROC indican un muy buen desempeño del modelo en la clasificación de las clases. Las clases A y A+ y D muestran un rendimiento perfecto, con un AUC de 1.00, lo que significa una capacidad perfecta para distinguir estas clases sin errores. Las clases B y C también

presentan buenos resultados, con AUC de 0.92, 0.89 respectivamente, aunque la clase C refleja mayor dificultad en comparación con las demás. Los promedios micro y macro, con valores de 0.97 y 0.96, confirman un alto rendimiento global del modelo, demostrando que maneja bien tanto las instancias individuales como el balance general entre las clases.

**Gráfica 5. Curva ROC de modelo ET**



Gráfica 6. Curva de validación de modelo ET



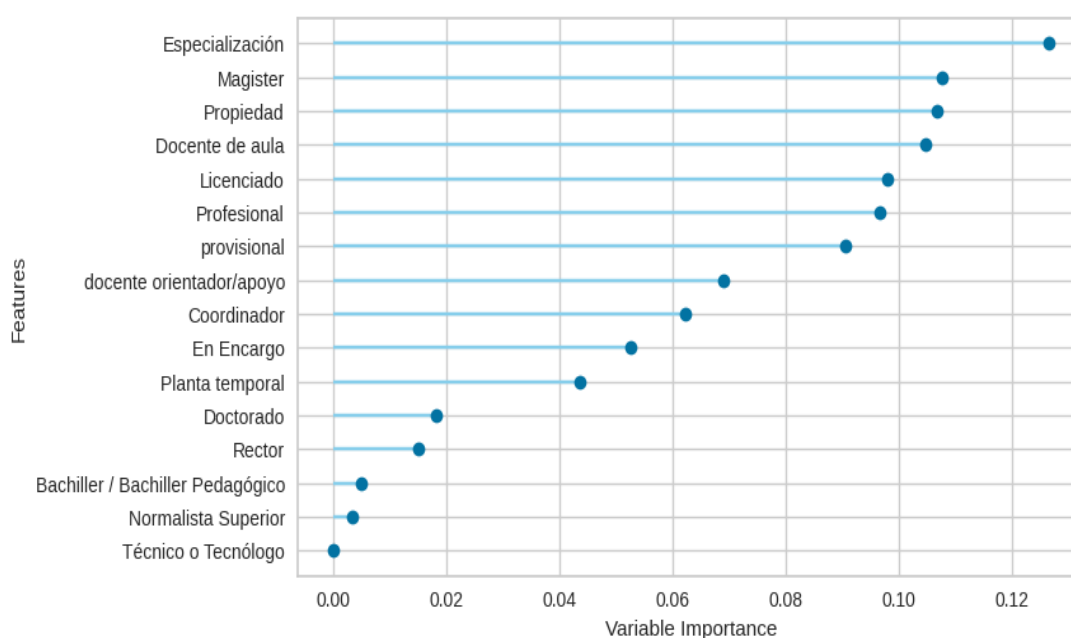
La curva de validación muestra el rendimiento del modelo en función del parámetro `max_depth`, que define la profundidad máxima de los árboles de decisión. A medida que `max_depth` aumenta, la precisión en los datos de entrenamiento se eleva gradualmente, alcanzando prácticamente el 100% alrededor de un `max_depth` de 8, lo que sugiere un fuerte sobreajuste, ya que el modelo memoriza los datos de entrenamiento. Sin embargo, la precisión en la validación cruzada, aunque también mejora a medida que aumenta `max_depth`, se estabiliza alrededor del 80% y presenta una banda de variabilidad, lo que indica que el modelo no generaliza tan bien a datos no vistos. Esto sugiere que limitar `max_depth` podría ser útil para reducir el sobreajuste y lograr un mejor equilibrio entre entrenamiento y validación.

En este sentido, la configuración de nuestro modelo que optimiza el rendimiento y evita el sobreajuste tiene los parámetros de `min_samples_split` en 7 y `min_samples_leaf` en 4 restringiendo la profundidad y la creación de nodos y reduciendo el riesgo de sobreajuste. Además, la profundidad máxima de los árboles (`max_depth`) está limitada a 6, lo cual ayuda a mantener la generalización del modelo.

### Importancia de las características

La gráfica 10 muestra las características que mayor influencia tienen en las predicciones del modelo ExtraTreesClassifier. La variable especialización aparece como la más importante, con un valor cercano a 0.13, lo que indica que los docentes con estudios de especialización son particularmente influyentes en los resultados del modelo. Le sigue Magister, con un valor de importancia cercano a 0.10, sugiriendo que el nivel de formación avanzada de los docentes es un factor clave en las predicciones. La tercera característica en importancia es "Propiedad", también con un valor cercano a 0.10, resaltando el impacto de la estabilidad laboral de los docentes en el desempeño del modelo.

### Gráfica 7. Importancia de características



Otras variables como Docente de aula, Licenciado, Profesional y provisional muestran una importancia moderada, con valores entre 0.06 y 0.08. Estas características también contribuyen significativamente a las predicciones del modelo. Docente de aula refleja la influencia del rol directo de los docentes en el aula, mientras que Licenciado y Profesional son importantes así no sea formación avanzada.

Finalmente, características como Coordinador, En Encargo, Planta temporal, Doctorado, y Rector tienen valores de importancia menores a 0.05, lo que sugiere que, en comparación con las variables principales, su influencia en el modelo es limitada.

## Interpretabilidad del Modelo

SHAP calcula el impacto marginal que tiene una característica en la predicción de cada instancia específica. SHAP deriva del concepto de valores de Shapley de la teoría de juegos, lo que garantiza una distribución justa de la contribución de cada característica. (SHAP, 2018)

Mientras que Feature Importance provee una visión global de cuáles son las características más influyentes en el modelo, SHAP proporciona más detalles, tanto a nivel global como local, explicando cómo cada característica afecta cada predicción específica y qué interacciones entre las características tienen lugar.

Es importante mencionar que, en su documentación, SHAP enfatiza en que sus valores hacen transparentes las correlaciones recogidas por los modelos predictivos de ML. “Pero hacer que las correlaciones sean transparentes no las convierte en causales”, así que usar modelos predictivos para guiar elecciones de política pública puede ser engañoso (SHAP, 2018).

**Gráfica 8. Importancia de Características Basada en Valores SHAP por Clase"**



La gráfica 11 muestra cómo las características específicas influyen en la predicción del modelo para cada clase de nuestro modelo. En la clase A, las características más influyentes son

Coordinador y Propiedad, lo que sugiere que los roles administrativos y la estabilidad laboral de los docentes juegan un papel importante en esta categoría superior. Además, cualificaciones como Magister y Especialización también contribuyen de manera significativa. Para la clase A+, las características de mayor impacto son Especialización y Profesional, pero también Magister y Docente de aula son importantes,

En la clase B, las características Docente de aula y Magister son las más influyentes, reflejando la importancia del rol en el aula y la formación avanzada para esta categoría intermedia. Otras características como Propiedad y Profesional también aportan a la predicción, aunque en menor medida.

Finalmente, en la clase D, predominan las características Propiedad y Docente de aula, En esta clase, la cualificación avanzada tiene una influencia menor en comparación con las clases superiores, aunque características como docente orientador/apoyo y Especialización siguen siendo relevantes.

## **Conclusiones y Discusión**

El análisis de datos sobre formación, vinculación y roles docentes en las instituciones educativas de Cundinamarca, combinado con modelos de aprendizaje automático, ha permitido identificar factores que influyen en el rendimiento académico de los estudiantes y, por ende, en la clasificación de los colegios. La aplicación de técnicas de machine learning, específicamente el modelo Extra Trees Classifier, mostró un rendimiento moderado en la predicción de las categorías de clasificación. No obstante, fue más eficaz al diferenciar entre clases, resaltando el valor de la ciencia de datos para analizar el impacto de las características docentes en los resultados educativos y como una herramienta potencial para fundamentar decisiones de políticas basadas en evidencia.

Aunque el modelo mostró un desempeño moderado, los resultados deben interpretarse con precaución, ya que el rendimiento en entrenamiento y validación interna no garantiza robustez en contextos reales. Es necesario validar el modelo en otros períodos y con datos de diferentes regiones para mejorar sus métricas y permitir una mayor generalización de las conclusiones.

Los resultados destacan la formación en especializaciones y maestrías, la vinculación en propiedad y el ejercicio directo en el aula como variables clave en las predicciones del modelo. En cambio, las variables relacionadas con los directivos docentes aportaron poco, lo cual

resulta decepcionante, dado que rectores y coordinadores, con alta formación, desempeñan funciones cruciales en la planificación y coordinación curricular de las instituciones educativas.

El análisis SHAP revela además que, para las clases superiores (A+ y A), la formación en maestría no es la característica más relevante: en la clase A+, la especialización y la formación universitaria ocupan un lugar de mayor importancia, mientras que, en la clase A, la posición de coordinador y la vinculación en propiedad son más determinantes. Curiosamente, en ninguna clase la formación doctoral aparece como una variable importante para el modelo.

En particular, el desbalance de datos, especialmente en las clases superiores A y A+, representó un desafío significativo. El uso de técnicas de balanceo mejoró la distribución de las clases, logrando una representación más equitativa en el conjunto de datos; sin embargo, estas técnicas pueden aumentar el riesgo de sobreajuste. En Colombia, la clase A+ representa solo el 1.66% de las IE oficiales, por lo que identificar los factores que contribuyen a su éxito es una línea de investigación importante para futuros estudios.

Es importante subrayar que la mejora en la calidad educativa es un fenómeno multicausal, en el cual el nivel de formación de los docentes es solo una variable dentro de un contexto más amplio que incluye factores administrativos, socioeconómicos y de gestión escolar. Así, aunque este estudio se enfoca en la formación, vinculación y rol docente, no sugiere que esta sea la mejor vía para elevar los resultados educativos. Más bien, estos hallazgos pueden servir de apoyo para políticas que integren formación avanzada de docentes con estrategias pedagógicas innovadoras y una gestión escolar efectiva.

Los resultados de este trabajo invitan a futuras investigaciones que profundicen en la relación entre la formación docente y el rendimiento académico, idealmente en un marco de políticas educativas que fomenten no solo el nivel académico de los docentes, sino también prácticas pedagógicas adaptativas y enfoques de gestión escolar que permitan maximizar el impacto positivo de la educación en los estudiantes.

## **Bibliografía**

- Alcaldía Mayor de Bogotá. (2012). *Plan de desarrollo 2012-2016 Bogotá Humana*. Bogotá.
- Alyahyan, E., & Düşteğör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational*

*Technology in Higher Education*. Obtenido de <https://doi.org/10.1186/s41239-020-0177-7>

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics*, 66, 103-115. Obtenido de <https://www.sciencedirect.com/science/article/abs/pii/S0094119009000291>

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 673-682.

Contreras, L., Tarazona, G., & y otros. (2023). *Machine Learning aplicado al rendimiento académico en la educación superior: factores, variables y herramientas*. Bogotá: UD Editorial.

Croninger, R., Rey Arroz, J., Rathbun, U., & Nishio, M. (s.f.). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement.

De Zubiría, J. (22 de septiembre de 2020). ¿Sirven las maestrías en educación? *El Espectador*.

DELIA, L. B. (2019). *IMPLICACIONES DEL AUSENTISMO LABORAL DE LOS DOCENTES EN UNA*. Bogotá: FUNDACIÓN UNIVERSIDAD DE AMÉRICA.

Gaviria, A., & Barrientos, J. H. (2001). *Fedesarrollo*. Obtenido de <https://www.repository.fedesarrollo.org.co/handle/11445/1249>

Hammed, I., Mahboob, T., & Mushtaq, M. (2022). A Novel Method for Performance Measurement of Public Educational Institutions Using Machine Learning Models. *Applied Sciences*.

Hanushek, E. (2003). The Failure of Input-Based Schooling Policies. *The Economic Journal*(485), 64-98. Obtenido de <https://www.jstor.org/stable/3590139>

<https://imbalanced-learn.org/>. (2024). Obtenido de [https://imbalanced-learn.org/stable/over\\_sampling.html#smote-adasyn](https://imbalanced-learn.org/stable/over_sampling.html#smote-adasyn)

ICFES. (2017). *Icfes*. Obtenido de <https://www2.icfes.gov.co/documents/39286/1648686/Clasificacion+de+establecimientos+y+sedes+Saber+11.pdf/6a5b0739-9cd2-d231-14e6-2a19df60a5b6?version=1.0&t=1647380417344#:~:text=La%20clasificaci%C3%B3n%20se%20realiza%20en%20cinco%20categor%C3%ADas%3A%20A%20>

ICFES. (agosto de 2023). ¿Qué es y como se genera la clasificación de los planteles educativos? *Saber al detalle*. Bogotá.

- Issah, I., & y otros. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision Analytics Journal*.
- Long, P., & Siemens, G. (12 de 09 de 2011). *Penetrating the Fog: Analytics in Learning and Education*. Obtenido de EDUCAUSE Review: <https://er.educause.edu/>
- Lucía, C. O. (2014). *De los bordes que delinear la formación docente: Planes territoriales, incentivos, necesidades, alternativas de cualificación docente ho*. Bogotá.
- MEN. (2008). *Plan Sectorial 2006-2010. La revolución Educativa*. Bogotá.
- MEN. (2011). *Plan Sectorial de Educación 2010-2014*. Bogotá.
- MEN. (2013). *SISTEMA COLOMBIANO DE FORMACIÓN DE EDUCADORES Y LINEAMIENTOS DE POLÍTICA*. Bogotá.
- MEN. (21 de Enero de 2022). *El Icfes trabaja por la calidad de la educación, a través de la evaluación*. Obtenido de <https://www.mineducacion.gov.co/portal/salaprensa/Comunicados/409033:El-Icfes-trabaja-por-la-calidad-de-la-educacion-a-traves-de-la-evaluacion#:~:text=Entre%20las%20Entidades%20Territoriales%20Certificadas%20con%20mayor%20poblaci%C3%B3n,permanente%20entre>
- MEN. (29 de Septiembre de 2023). El presidente Gustavo Petro anunció 25 mil becas para maestras y maestros de Colombia del Programa Poder Pedagógico. Bogotá.
- Miller, R., Murnane, R., & Willett, J. (2008). Do Teacher Absences Impact Student Achievement? Longitudinal Evidence from One. *Educational Evaluation and Policy Analysis*, 181-200.
- Ortega, L., Malmberg, E., & Sammon, P. (2018). Teacher Effects on Chilean Children's Achievement Growth: A Cross-Classified Multiple Membership Accelerated Growth Curve Model. *Educational Evaluation and Policy Analysis*, 473-501.
- Pycaret.org. (04 de 2020). *PyCaret*. Obtenido de <https://pycaret.org/>
- Rivkin, S., & Schiman, J. (2015). Instruction time, Classroom Quality, and Academic Achievement. *The Economic Journal*.
- SED. (2018). *Política Pública para la Formación de Docentes y directivos docentes en el distrito capital*. Bogotá.
- SHAP. (2018). Obtenido de <https://shap.readthedocs.io/en/latest/index.html>
- SIMAT. (s.f.). 2023.
- Staiger, D., & Rockoff, J. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*.

Zaman, B., Sharma, A., & Ram, C. (2023). Modeling education impact: a machine learning-based approach for improving the quality of school education. *Journal of computer in education*.