Proyecto Institucional

Aplicación de Machine Learning para justificar el agrupamiento por niveles de rendimiento en 7º Básico

Autor: Miguel San Juan

Fecha: 8 de junio de 2025

Este proyecto se presenta como propuesta de innovación pedagógica basada en inteligencia artificial educativa.

Índice

1 Identidad y Objetivos del Proyecto

Este proyecto presenta una propuesta innovadora en el ámbito pedagógico, fundamentada en la ciencia de datos y enfocada en fortalecer la toma de decisiones educativas mediante la aplicación de inteligencia artificial. Surge específicamente de la necesidad del Instituto Regional del Maule, ubicado en la comuna de San Javier, de contar con evidencia objetiva que respalde la separación de estudiantes de 7º básico según sus niveles de desempeño académico.

En esta primera etapa del proyecto, que se detalla en este informe, el propósito es desarrollar un modelo supervisado de clasificación basado en técnicas avanzadas de aprendizaje automático. Este modelo permitirá predecir con precisión el nivel académico (Avanzado o No Avanzado) que tendrán los estudiantes al ingresar a 7º básico, utilizando para ello datos históricos sobre calificaciones y asistencia desde 1º hasta 6º básico. De esta forma, se busca proporcionar fundamentos empíricos robustos que apoyen la decisión de agrupar a los estudiantes según su rendimiento escolar.

2 Introducción

El sistema educativo actual enfrenta desafíos importantes relacionados con la necesidad de acoger y responder de manera efectiva a la creciente diversidad de estudiantil. Esta diversidad se manifiesta a través de múltiples factores, tales como: trayectorias escolares diferenciadas, estilos de aprendizaje variados, contextos socioculturales diversos, distintos niveles de apoyo familiar y realidades emocionales particulares, configurando un escenario complejo donde la enseñanza estandarizada y homogénea resulta insuficiente para garantizar el desarrollo integral de todos los estudiantes.
Uno de los principales efectos de esta diversidad es la manifestación de distintos ritmos y formas de aprendizaje en el aula. Mientras algunos estudiantes avanzan rápidamente en la comprensión de contenidos, otros requieren más tiempo, apoyo diferenciado y estrategias específicas que les permitan progresar adecuadamente en su aprendizaje. Esta situación exige que los equipos docentes diseñen e implementen propuestas pedagógicas inclusivas, flexibles y basadas en evidencia, que respondan a las necesidades reales de los estudiantes y promuevan trayectorias educativas significativas para cada uno de ellos.
En este contexto, algunas instituciones han optado por estrategias de diferenciación estructural, como la separación de estudiantes por niveles de desempeño académico. Esta práctica, ha sido adoptada en distintas formas con el objetivo de adecuar la enseñanza al nivel de los estudiantes, buscando mejorar su rendimiento. Sin embargo, en nuestro país, dicha medida ha suscitado una serie de tensiones y cuestionamientos pedagógicos, éticos y sociales. Diversos estudios y debates han señalado que, si bien puede facilitar la planificación docente en el corto plazo, también puede reforzar estigmas, limitar las oportunidades de aprendizaje de ciertos grupos y reproducir inequidades al interior del sistema escolar.
Debido a que esta práctica se realiza en el establecimiento y dado el impacto que puede tener en la trayectoria educativa de los estudiantes, es importante someterla a un proceso riguroso de evaluación y validación, que permita comprender sus efectos reales, sus fundamentos pedagógicos y las oportunidades de mejora. En particular, se plantea el uso de herramientas de análisis de datos y modelos predictivos que permitan observar patrones de desempeño académico a lo largo del tiempo y anticipar posibles escenarios de aprendizaje, contribuyendo así a la toma de decisiones más informada y contextualizada.
En este marco, la presente proyecto busca aplicar técnicas de Machine Learning sobre datos académicos históricos de estudiantes desde 1º a 6º básico, con el objetivo de predecir el nivel de desempeño con el que ingresarán a 7º básico y, a partir de ello, evaluar empíricamente la pertinencia del agrupamiento por niveles. El propósito final es fundamentar esta práctica desde una perspectiva pedagógica crítica y proponer ajustes o alternativas que fortalezcan la equidad y calidad del proceso educativo.

2.1 Objetivo general

Implementar un modelo predictivo supervisado capaz de anticipar el nivel de desempeño académico de los estudiantes al ingresar a 7º básico, con el propósito de validar empíricamente la separación por niveles, explicarla pedagógicamente y optimizar sus criterios actuales.

2.2 Objetivos específicos

1. Recolectar, organizar y sistematizar datos académicos desde 1º a 6º básico.
2. Entrenar un modelo de *Machine Learning* supervisado que prediga el nivel de desempeño en 7º básico.
3. Comparar la predicción del modelo con la asignación histórica de niveles para evaluar la coherencia del sistema.
4. Generar visualizaciones e interpretaciones pedagógicas claras de los resultados del modelo.
5. Formular recomendaciones prácticas para mejorar los criterios actuales de agrupamiento por niveles.

3 Descripción del Conjunto de Datos

3.1 Fuentes de datos

Los datos utilizados en este proyecto provienen de los registros académicos históricos del Instituto Regional del Maule, establecimiento educacional ubicado en la comuna de San Javier. La información considerada corresponde a calificaciones finales anuales por asignatura, registradas desde 1º hasta 7º básico, además de los porcentajes de asistencia anual de cada estudiante. Estos datos fueron recopilados y organizados con el propósito de analizar trayectorias escolares longitudinales y desarrollar un modelo predictivo basado en evidencia objetiva.

Además, los datos incluyen la asignación del nivel académico, *"Avanzado"* o *"No Avanzado"*, que cada estudiante recibió al ingresar al 7º básico. Este nivel fue determinado históricamente por la institución con base en criterios internos definidos por el equipo docente y la dirección académica, fundamentalmente relacionados con los promedios generales de las asignaturas científico-humanistas obtenidos por los estudiantes en sexto básico.

A pesar de este criterio específico, se optó por entrenar al modelo con las notas obtenidas durante toda su trayectoria académica, desde 1º hasta 6º básico, con la finalidad explícita de identificar diferencias o similitudes relevantes. Esto permite realizar una evaluación empírica profunda del criterio previamente establecido, buscando determinar si dicho criterio captura efectivamente la evolución académica de los estudiantes o si existen patrones adicionales que podrían contribuir a mejorar la precisión y equidad del agrupamiento académico.

En un inicio, se proyectó trabajar con cinco generaciones de estudiantes. Sin embargo, debido a inconsistencias en la calidad y completitud de los datos entregados, fue necesario restringir el análisis a tres cohortes: Generación 2017, 2018 y 2019. Las generaciones 2020 y 2021 fueron descartadas, ya que, producto del contexto de emergencia sanitaria asociado a la pandemia, no se aplicó la diferenciación por niveles de ingreso en 7º básico, lo que impedía su utilización como datos etiquetados para un modelo supervisado.

Otro aspecto relevante es que los registros originales fueron entregados por la institución en formato PDF, lo que implicó una etapa adicional de extracción, depuración y estructuración. Para ello, se utilizó la herramienta Power Query de Microsoft Excel, lo que

permitió transformar múltiples tablas provenientes de distintos archivos en un único conjunto de datos limpio, estructurado y consolidado en formato CSV, apto para el análisis con herramientas de machine learning.
La obtención y uso de estos datos se realizó con autorización explícita del Instituto Regional del Maule, con el propósito específico de desarrollar análisis pedagógicos e investigaciones aplicadas destinadas a mejorar la gestión académica y apoyar decisiones basadas en evidencia.

3.2 Tratamiento ético y resguardo de información

- Todos los datos fueron completamente anonimizados, eliminando cualquier identificador personal que pudiera asociarse a estudiantes específicos.
- La información fue utilizada exclusivamente con fines pedagógicos y para apoyar procesos de mejora institucional basados en evidencia.
- No se realizó ningún contacto directo con los estudiantes, ni se aplicaron instrumentos adicionales durante el desarrollo del proyecto.
- Se aseguró el cumplimiento riguroso de la Ley N.º 19.628 sobre Protección de la Vida Privada, resguardando la confidencialidad y el uso ético de los datos escolares.

3.3 Distribución de los datos

El análisis exploratorio de datos tuvo como objetivo comprender la estructura interna del dataset, identificar patrones relevantes, detectar posibles valores atípicos y evaluar la calidad de las variables empleadas para la predicción del nivel en 7º básico. Esta etapa permitió obtener evidencia empírica que permite entregar fundamentos sobre la elección del modelo, confirmar su aplicabilidad y analizar pedagógicamente la coherencia de los agrupamientos actuales.

3.3.1 Distribución de la variable objetivo

La variable objetivo del modelo se denomina Nivel 7, esta representa la categoría en la que fue clasificado cada estudiante al ingresar a 7º básico: *Avanzado* o *No Avanzado*. Dado que se trata de un problema de clasificación supervisada, la distribución de clases es un aspecto relevante para evaluar la calidad del dataset y la estrategia de partición de datos para el entrenamiento y la prueba.
La Figura 1 muestra un gráfico de barras con la frecuencia de estudiantes por Nivel. Se observa que:

- 98 estudiantes fueron clasificados como *No Avanzado*.
- 95 estudiantes como *Avanzado*.

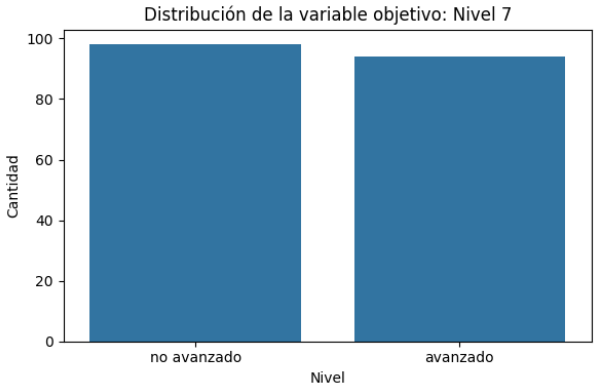

Figura 1: Distribución de la variable objetivo: Nivel en séptimo básico

Esta diferencia mínima refleja una distribución casi balanceada entre ambas clases, lo cual es ideal para entrenar un modelo de clasificación binaria. En particular:

- No hay dominancia de una clase sobre la otra, lo que evita que el modelo aprenda un sesgo hacia la clase mayoritaria.
- Se justifica el uso de *Accuracy* y *F1-score* como métricas principales, ya que la distribución no requiere técnicas de balanceo o penalización por desbalance.

Desde una perspectiva pedagógica, esta distribución sugiere que la institución ha intentado mantener una proporción equilibrada en la asignación de estudiantes por nivel, lo que permite evaluar con mayor objetividad la calidad del proceso de agrupamiento y su justificación empírica a través del modelo.
Además, esta representación visual cumple una función importante en el análisis exploratorio, ya que nos sugiere que el conjunto de datos está bien constituido para tareas de clasificación y que no requiere técnicas avanzadas de remuestreo para su entrenamiento inicial.

3.3.2 Distribución de promedios generales y asistencia

Se analizaron histogramas y diagramas de caja para las siguientes variables predictoras:

- Promedios de cada asignatura desde 1º a 6º básico.
- Promedio general de asistencia.

Los datos analizados revelaron los siguientes patrones:

- Una concentración de notas en el rango de 5.5 a 6.3 aproximadamente, con distribuciones, por lo general, asimétricas hacia la izquierda, lo que indica que la mayoría de los estudiantes presenta calificaciones altas, aunque existen algunos casos con notas considerablemente más bajas que el promedio.
- Una asistencia promedio centrada en torno al 95 %, aunque con una cola hacia valores bajos, con mínimos cercanos al 70 %.Esto evidencia la presencia de casos aislados de inasistencia significativa, que podrían tener impacto en el rendimiento académico.

Estos patrones son consistentes con un sistema educativo que mantiene un estándar académico medio-alto en general, pero que también presenta ciertos casos críticos de asistencia reducida, los cuales podrían correlacionarse con un menor rendimiento académico.

3.3.3 Distribución de las calificaciones de Matemáticas y Lenguaje: comparación entre 6º y 7º básico

Para observar la evolución del rendimiento académico en Matemática y Lenguaje a lo largo del tiempo, se construyeron histogramas correspondientes a las notas finales desde 1º a 7º básico. El objetivo de este análisis fue identificar posibles cambios significativos en el desempeño de los estudiantes, especialmente en 7º básico, tras la implementación de la categorización por niveles.
Estos gráficos permiten examinar no solo la frecuencia relativa de los distintos niveles de logro, sino también la forma de las distribuciones, incluyendo aspectos como la simetría, la concentración y el sesgo, lo que aporta evidencias relevantes para el análisis pedagógico y la toma de decisiones educativas.

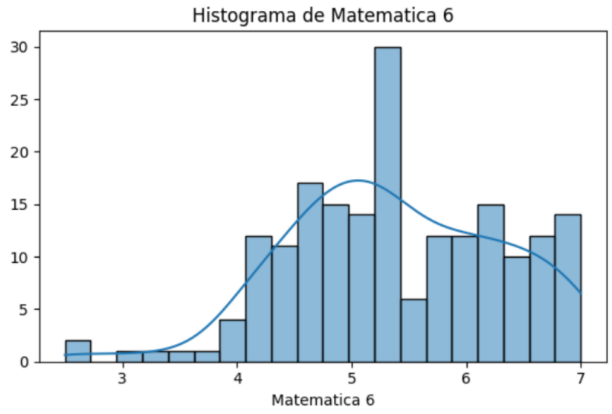

Figura 2: Histograma de Matemática 6º básico

Este gráfico representa la frecuencia de las calificaciones obtenidas por los estudiantes en 6º básico, con una curva de densidad suavizada superpuesta para facilitar la interpretación. En él se observa lo siguiente:

- La distribución muestra una leve asimetría hacia la izquierda. Existe una mayor concentración de estudiantes en los tramos de calificaciones entre 4.5 y 6.0 aproximadamente, aunque también se detectan picos irregulares, especialmente en torno al 5.0 y entre 6.5 y 7.0.
- Se identifican algunos casos extremos con notas cercanas a 2.5, aunque su frecuencia es baja.
- El grupo más numeroso se ubica en torno al promedio aritmético 5.0, pero la abundancia de estudiantes en tramos inferiores evidencia una alta dispersión en el rendimiento académico durante esta etapa.

Desde el punto de vista pedagógico, la dispersión en los resultados sugiere una creciente heterogeneidad en el dominio de las habilidades matemáticas a medida que los estudiantes avanzan hacia los niveles intermedios de la educación primaria. Esta etapa suele marcar el inicio de una diferenciación más notoria en el rendimiento académico, con algunos alumnos consolidando aprendizajes y otros comenzando a presentar dificultades sostenidas.
Por otro lado, para el caso de séptimo básico, se observa lo siguiente:

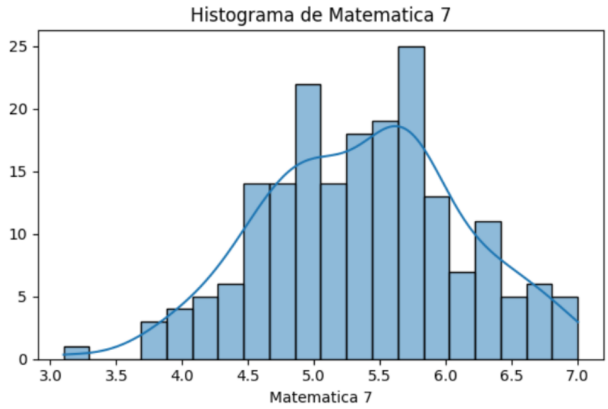

Figura 3: Histograma de Matemáticas 7° básico

Este gráfico representa las calificaciones obtenidas por los mismos estudiantes al finalizar 7º básico, permitiendo comparar directamente la evolución de la distribución respecto del año anterior.

- La distribución es más simétrica con una forma que se aproxima a la distribución normal, centrada en torno al 5.3 y 5.5 aproximadamente.
- Hay una disminución de casos extremos bajos y también una reducción de la cantidad de estudiantes en los tramos superiores.
- Esto indica una mayor concentración de los estudiantes en el rango medio de desempeño, lo que puede reflejar un proceso de nivelación o estabilización pedagógica.

Este resultado es relevante porque sugiere que, tras el agrupamiento por niveles y la intervención docente en 7º básico, la dispersión se reduce y la distribución se normaliza, lo que puede ser indicativo de un efecto pedagógico correctivo o regulador del sistema de categorización por niveles.

Implicancias del análisis de histogramas

La comparación entre ambas histogramas permite concluir lo siguiente:

- En 6º básico, la variabilidad del rendimiento es mayor, con diferencias marcadas entre estudiantes de alto y bajo desempeño.
- En 7º básico, la distribución se vuelve más concentrada, lo que sugiere que la acción educativa implementada podría estar ayudando a reducir las brechas extremas.
- Este cambio estructural en la forma de la distribución respalda lo observado en los diagramas de caja, la brecha entre los grupos disminuye, y el sistema parece avanzar hacia una mayor equidad interna.

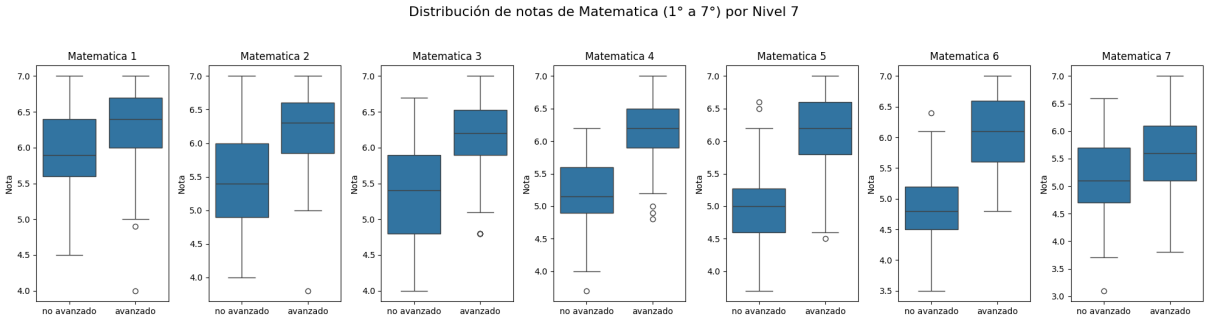

Figura 4: Diagramas de caja de Matemáticas

Desde el punto de vista del modelo, esto también implica que el desempeño de 6º básico contiene más variabilidad explicativa y es más útil para predecir la pertenencia a un grupo en 7º. En cambio, el rendimiento de 7º básico podría reflejar más los efectos del contexto o de la intervención pedagógica, y no tanto los factores históricos previos.

En el caso de **lenguaje**, al realizar un análisis análogo al anterior, la comparación entre los histogramas de 6º y 7º básico permite concluir lo siguiente:

- El rendimiento en Lenguaje se estabiliza al ingresar a 7º básico, con una reducción en la dispersión y una mayor concentración en torno al promedio.
- Se produce un efecto correctivo similar al observado en Matemáticas, disminuyen los extremos, se modera la variabilidad y se consolida el nivel medio.
- Esto sugiere que la estrategia de categorización podría estar contribuyendo a equilibrar las trayectorias académicas y a reducir las brechas de aprendizaje dentro del curso.

Además, al observar los diagramas de caja, se puede evidenciar que cada gráfico representa de manera clara la mediana, el rango intercuartil, la dispersión total de las calificaciones y la presencia de posibles valores atípicos. El análisis secuencial desde 1º hasta 7º básico revela patrones consistentes y sostenidos de diferencia en el rendimiento académico entre los grupos Avanzado y No Avanzado. Esta evolución progresiva de las trayectorias escolares permite evaluar con mayor fundamento la validez del proceso de categorización aplicado en 7º básico, así como comprender el desarrollo previo de los aprendizajes en la asignatura.

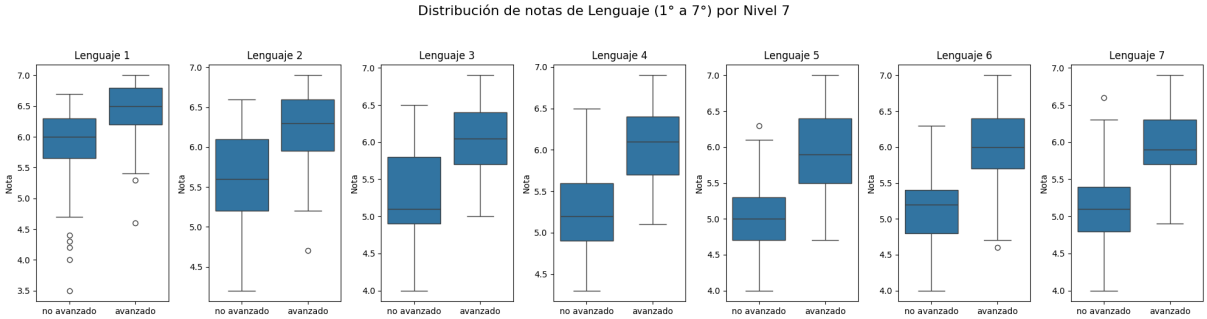

Figura 5: Diagramas de caja de Lenguaje

Al observar las calificaciones en 7º básico, se percibe una estabilización en ambas trayectorias: la mediana del grupo *Avanzado* se mantiene alta, alrededor de 6.1, mientras que la del grupo *No Avanzado* aumenta levemente hacia 5.2, con una dispersión menor en comparación con los niveles anteriores. Esta reducción en la variabilidad puede interpretarse, desde una perspectiva pedagógica, como un efecto positivo de la categorización por niveles y de las intervenciones aplicadas, las cuales habrían contribuido a contener y acompañar especialmente al grupo con menor rendimiento previo. Sin embargo, desde el punto de vista del modelo predictivo, esta misma estabilización tiene una implicancia técnica importante: al homogeneizarse las calificaciones en 7º básico, las notas de ese nivel reflejan en mayor medida el impacto de la intervención pedagógica que el rendimiento histórico puro, lo que disminuye su capacidad discriminante para predecir la pertenencia al grupo *Avanzado* o *No Avanzado*. Por esta razón, las calificaciones de 6º básico resultan ser más informativas para el modelo, ya que representan de forma más directa la trayectoria previa sin el efecto regulador del proceso de nivelación.

3.3.4 Análisis de Correlación entre Rendimiento Histórico y Nivel en 7º Básico

Con el propósito de identificar la relación entre el desempeño académico previo y la categorización de los estudiantes al ingresar a 7º básico, se calcularon coeficientes de correlación de Pearson entre los promedios anuales globales desde 1º a 6º básico y la variable objetivo codificada como 1 para aquellos estudiantes de nivel *avanzado* y 0 para los *no Avanzado*.
Estas relaciones se representaron visualmente mediante matrices de calor, lo que permite observar con claridad la intensidad y dirección de las correlaciones.

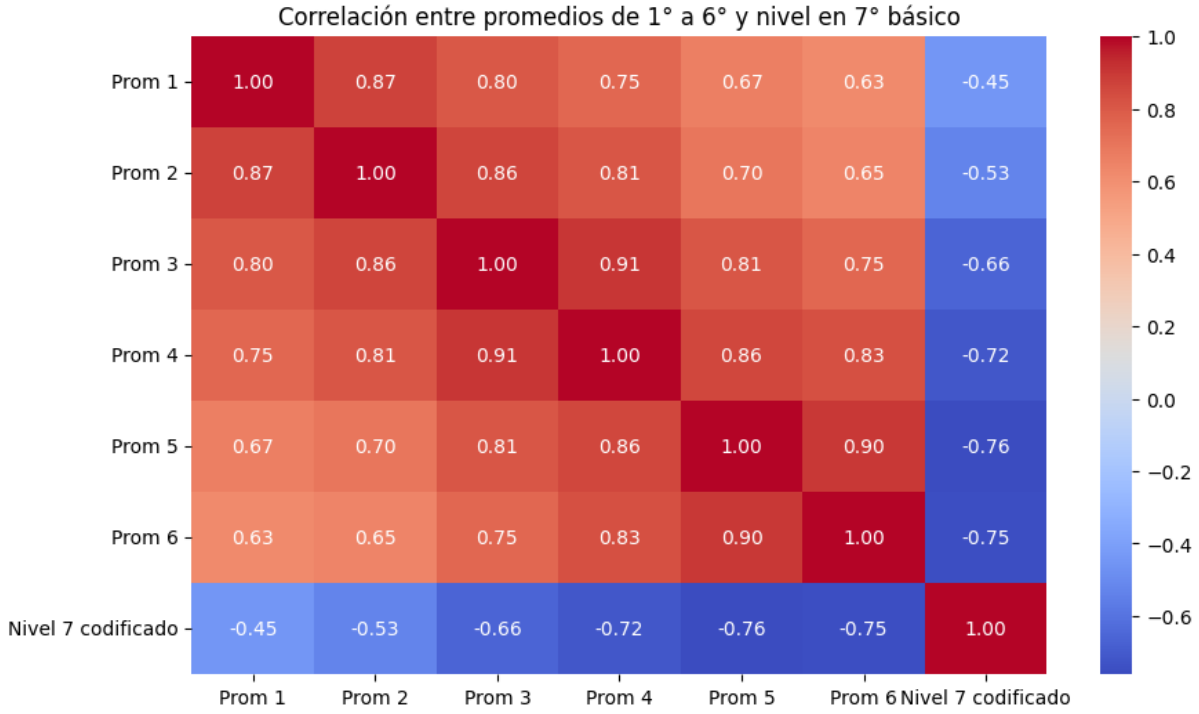

Figura 6: Matriz de correlación entre promedios de 1º a 6º básico y nivel en 7º básico.

La Figura 6 muestra la matriz de correlación entre los promedios anuales generales de cada curso y la variable de nivel. Se observa una alta consistencia longitudinal en el rendimiento académico: las correlaciones entre los promedios de distintos años varían entre 0.63 y 0.91. En cuanto a la correlación entre los promedios y el nivel en 7º básico, los valores son positivos, lo que refleja que los estudiantes con promedios más altos tienen mayor probabilidad de haber sido clasificados como *Avanzados*. Esta relación se vuelve más fuerte con el paso de los años escolares: desde r = 0.45 en Promedio 1 hasta r = 0.76 en Promedio 6. Desde una perspectiva pedagógica, este comportamiento indica que el rendimiento escolar es acumulativo y predictivo. En particular, los promedios de 5º y 6º básico resultan ser los mejores indicadores del nivel de rendimiento futuro, lo cual valida su uso como variables clave en el modelo de clasificación y para orientar decisiones institucionales sobre agrupamientos o apoyos diferenciados.

3.3.5 Preprocesamiento de los datos

El preprocesamiento fue una etapa ardua, pero fundamental para garantizar la calidad del conjunto de datos y preparar adecuadamente las variables antes de aplicar los modelos de aprendizaje. A continuación, se detallan los principales pasos realizados:

- **Construcción inicial del dataset:** El conjunto de datos fue construido utilizando Power Query, debido a que la información original proporcionada por el establecimiento educacional se encontraba en archivos PDF. Durante el proceso de extracción y estructuración de los datos, se identificaron diversas inconsistencias en los registros, lo que impactó en la calidad y cantidad de información disponible para el análisis.
- **Copia de trabajo:** Se creó una copia del dataset original (`df_filtrado`) para preservar los datos sin modificaciones y realizar transformaciones de forma segura y controlada.
- **Cálculo de valores nulos:** Se añadió una columna auxiliar llamada `nulos`, que contabiliza cuántos valores faltantes presenta cada fila.
- **Eliminación de registros incompletos:** Se eliminaron todos los registros con más de 30 columnas nulas, manteniendo solo estudiantes con suficiente información para el análisis.
- **Corrección de formato numérico:** Las columnas de promedios anuales (`Prom 1` a `Prom 6`) contenían valores en formato texto con comas como separador decimal. Se reemplazaron las comas por puntos y se convirtieron los datos en float.
- **Cálculo del promedio general del estudiante:** Se calculó el promedio de los promedios anuales, generando una nueva variable llamada *Promedio general estudiante* que resume el rendimiento histórico del alumno.
- **Imputación de valores faltantes:** Se imputaron los valores faltantes en variables numéricas utilizando el *Promedio general estudiante*. Esto permitió mantener coherencia para cada estudiante y evitar eliminar más registros.
- **Eliminación de filas sin variable objetivo:** Se eliminaron todas las observaciones en las que la variable Nivel 7 estaba vacía, ya que no pueden usarse para entrenamiento supervisado.

- **Codificación de la variable objetivo:** Se transformó la variable categórica Nivel 7 en una variable binaria denominada Nivel 7 codificado:
  - `Avanzado` → 1
  - `No Avanzado` → 0
- **Selección de características predictoras:** Se excluyeron columnas irrelevantes como identificadores, fechas, variables auxiliares y la variable objetivo. Se conservaron solo las variables numéricas asociadas al rendimiento y asistencia de 1º a 6º básico.
- **Normalización:** Se aplicó `StandardScaler` para escalar las variables predictoras, de modo que todas tengan media cero y desviación estándar uno. Esto evita que variables con distintas escalas dominen el entrenamiento.
- **Reincorporación de la variable objetivo:** Una vez escaladas las variables predictoras, se agregó nuevamente la columna Nivel 7 codificado al conjunto de datos final.
- **Prevención de fuga de información:** Se eliminaron todas las columnas relacionadas con notas de 7º básico.
- **Resultado final:** El conjunto `df_final` contiene observaciones completas desde 1º a 6º básico, con variables predictoras numéricas normalizadas y la variable objetivo codificada.

4 Métodos y Modelos Utilizados

Con el propósito de predecir el nivel académico de los estudiantes al ingresar a 7º básico, el problema se considero como una tarea de **clasificación supervisada binaria**. Para abordarlo se entrenaron y evaluaron cinco modelos de aprendizaje, seleccionados por representar distintos enfoques metodológicos. En esta sección se detallan las características principales de cada modelo y el procedimiento empleado para evaluar su desempeño.
Para llevar a cabo la modelación se definieron las variables predictoras $\mathbf{X}$ como todas las columnas, diferentes a la variable objetivo $\mathbf{Y}$. Posteriormente, se dividió el conjunto de datos en dos subconjuntos mediante la función *train_test_split*:

- Entrenamiento 80 %
- Prueba 20 %

considerando estratificación por clases, para asegur así una distribución equilibrada de ambas categorías en cada subconjunto.

4.1 Modelos de Clasificación Utilizados

Se implementaron los siguientes cinco modelos de clasificación:

- **Random Forest Classifier**: Método de ensamblado basado en múltiples árboles de decisión, entrenados de manera aleatoria y en paralelo. La clase final se determina mediante votación mayoritaria entre todos los árboles.

  Ventajas: robustez frente al sobreajuste, manejo eficiente de variables categóricas y numéricas, y facilidad interpretativa parcial.

- **XGBoost (Extreme Gradient Boosting)**: Algoritmo basado en árboles secuenciales, donde cada árbol busca corregir los errores del modelo anterior. Destaca por su regularización incorporada, manejo eficiente de valores faltantes y optimización computacional.

  Ventajas: alta precisión, eficiencia computacional y resistencia al sobreajuste. .

- **Regresión Logística**: Modelo lineal que estima probabilidades de pertenencia a cada clase mediante una función logística. Es útil como modelo base y ofrece una alta interpretabilidad.

  Ventajas: interpretabilidad y eficiencia computacional.

  Limitaciones: dificultad para capturar relaciones no lineales sin transformaciones previas.

- **K-Nearest Neighbors (KNN)**: Modelo no paramétrico que clasifica una observación según la mayoría de las clases entre sus k vecinos más cercanos en el espacio de características.

  Ventajas: simplicidad y adaptación local efectiva.

  Limitaciones: sensibilidad a la escala y a la dimensionalidad de los datos.

- **Perceptrón Multicapa (MLPClassifier)**: Red neuronal artificial con estructura de capas interconectadas. Se utilizó una red con una capa oculta de 100 neuronas, con función de activación ReLU y optimizador Adam.

  Ventajas: capacidad de modelar relaciones complejas y no lineales.

4.1.1 Evaluación y Desempeño

Para cada modelo se calcularon las siguientes métricas:

- **Exactitud (Accuracy)**: proporción global de clasificaciones correctas.
- **Matriz de Confusión**: permite visualizar verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.
- **Reporte de Clasificación**: incluye *precision*, *recall* y *f1-score*, tanto por clase como promediado ponderadamente (*weighted avg*).

A continuación, se presenta una tabla con los valores de *precision*, *recall*, *f1-score* ponderado y *accuracy* para cada modelo:

Modelo	Precisión	Recall	F1-Score	Accuracy
Random Forest	0.90	0.90	0.90	0.90
XGBoost	0.88	0.87	0.87	0.87
Logistic Regression	0.84	0.84	0.84	0.84
KNN	0.84	0.84	0.84	0.84
MLP	0.90	0.90	0.90	0.90

Cuadro 1: Comparación de desempeño de modelos clasificatorios.

4.2 Análisis comparativo y selección del modelo

La Tabla 1 presenta el desempeño de cinco modelos de clasificación evaluados en términos de *precision*, *recall*, *f1-score* y *accuracy*. Estas métricas permiten evaluar la capacidad predictiva de cada algoritmo en la tarea de clasificar a los estudiantes en nivel *Avanzado* o *No Avanzado* al ingreso a 7º básico.

- **Precisión (Precision):** proporción de predicciones positivas que fueron correctas.
- **Exhaustividad (Recall):** proporción de casos positivos reales que fueron identificados correctamente.
- **F1-score:** media armónica entre precisión y recall, útil cuando las clases están desbalanceadas.
- **Exactitud (Accuracy):** proporción total de predicciones correctas.

Los resultados permiten identificar dos modelos con rendimiento superior: **Random Forest** y **MLP**, ambos con valores de 0.90 en todas las métricas evaluadas. En segundo lugar se ubica **XGBoost**, con métricas en torno al 0.87. Finalmente, los modelos **Logistic Regression** y **KNN** alcanzan un desempeño moderado, con valores de 0.84 en todas las métricas.
Aunque **MLP** y **Random Forest** muestran desempeños equivalentes, se justifica la elección del modelo **Random Forest** como opción final debido a las siguientes razones:

- **Robustez:** menor riesgo de sobreajuste gracias al uso de múltiples árboles de decisión.
- **Interpretabilidad:** permite identificar la importancia relativa de cada variable predictora.
- **Facilidad de implementación:** no requiere una arquitectura compleja ni ajustes avanzados de hiperparámetros.

En conclusión, el modelo **Random Forest Classifier** representa la mejor alternativa para el problema abordado, al combinar un alto rendimiento predictivo con una interpretación clara y una implementación sencilla, aspectos especialmente relevantes en contextos educativos y de toma de decisiones institucionales.

4.3 Random Forest Classifier

El modelo seleccionado para la tarea de predicción del nivel de ingreso de los estudiantes a 7º básico fue *Random Forest Classifier*, un algoritmo de ensamblado basado en múltiples árboles de decisión. Este método se caracteriza por combinar los resultados de múltiples árboles entrenados sobre subconjuntos aleatorios del conjunto de datos, reduciendo así la varianza del modelo y mejorando su capacidad de generalización.
Cada árbol de decisión es entrenado utilizando una muestra bootstrap del conjunto de entrenamiento, y en cada nodo del árbol se selecciona aleatoriamente un subconjunto de características para decidir la partición, lo cual introduce diversidad entre los árboles. La predicción final del modelo corresponde a la clase seleccionada por mayoría de votos entre todos los árboles del bosque.

4.3.1 Parámetros del Modelo

El modelo fue implementado utilizando la clase *RandomForestClassifier* de la biblioteca *scikit-learn*. Los principales hiperparámetros configurados fueron los siguientes:

- `n_estimators = 100`: número de árboles del bosque. Este valor permite un equilibrio entre desempeño y tiempo de cómputo.
- `random_state = 42`: semilla aleatoria utilizada para garantizar la reproducibilidad de los resultados.
- `criterion = 'gini'` (valor por defecto): métrica de impureza utilizada para dividir los nodos en cada árbol.
- `max_features = 'sqrt'` (valor por defecto): estrategia para seleccionar el número de características en cada división.

Estos valores se seleccionaron tras una exploración inicial de los resultados y en función de las buenas prácticas estándar para clasificación binaria sobre conjuntos de datos estructurados y normalizados.

5 Evaluación del Rendimiento del Modelo

Durante el desarrollo del modelo, se dividió el conjunto de datos de la siguiente forma:

- **Entrenamiento (80 %):** se utilizó para ajustar los parámetros del modelo y entrenar los árboles de decisión.
- **Prueba (20 %):** se reservó exclusivamente para evaluar el desempeño del modelo final sobre datos no vistos.

La partición se realizó utilizando la función *train_test_split*, con la opción *stratify = y*, lo que asegura que la proporción de clases *Avanzado* y *No Avanzado*, se mantenga equilibrada tanto en el conjunto de entrenamiento como en el de prueba. Esta separación permite validar la capacidad del modelo para generalizar su aprendizaje a nuevos datos.
El modelo Random Forest fue entrenado con los datos de entrenamiento y luego evaluado sobre los datos de prueba. Al comparar el desempeño en ambos conjuntos, se observó una

consistencia en las métricas obtenidas: accuracy, precisión, recall y F1-score, lo que indica que el modelo no sufre de sobreajuste ni de subajust. En otras palabras, el modelo logró aprender patrones relevantes sin memorizar los datos de entrenamiento y se desempeñó de forma efectiva sobre datos no vistos.

5.1 Métricas de Evaluación

Las métricas de evaluación son fundamentales en aprendizaje automático, ya que permiten cuantificar la calidad de las predicciones realizadas por un modelo. Sin estas métricas, no sería posible evaluar de manera objetiva si el modelo está funcionando correctamente ni compararlo con otros modelos alternativos.
En tareas de clasificación, como la desarrollada en este proyecto, es común utilizar las siguientes métricas:

- **Exactitud (Accuracy):** mide la proporción de predicciones correctas sobre el total de casos. Es útil cuando las clases están equilibradas.
- **Precisión (Precision):** indica qué porcentaje de las instancias clasificadas como positivas realmente lo eran. Es importante en contextos donde los falsos positivos tienen un alto costo.
- **Exhaustividad (Recall):** mide la proporción de verdaderos positivos que fueron correctamente identificados por el modelo. Es fundamental cuando el objetivo es no omitir casos positivos reales.
- **F1-Score:** media armónica entre precisión y recall, útil para evaluar el balance entre ambas.

Adicionalmente, se construyó una matriz de confusió para analizar los errores de clasificación, distinguiendo entre verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

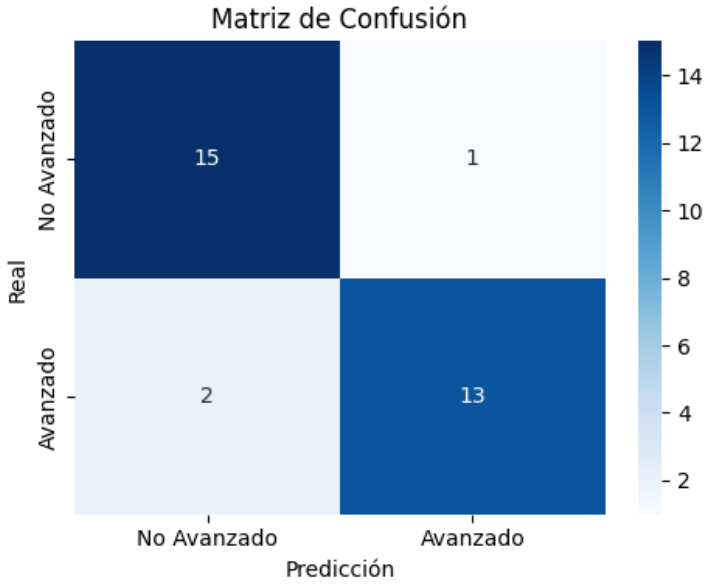

Figura 7: Matriz de Confusion

Los resultados del modelo mostrados en la Tabla 1 indican un desempeño equilibrado y robusto del modelo, con una alta capacidad para identificar correctamente tanto a los estudiantes *Avanzados* como a los *No Avanzados*. La matriz de confusión confirma que el número de errores fue bajo y distribuido simétricamente entre las clases.
Ademas, el modelo Random Forest no solo obtuvo las mejores métricas entre los modelos evaluados, sino que también ofrece ventajas metodológicas relevantes: es menos susceptible al sobreajuste, permite el análisis de importancia de variables y su comportamiento es consistente ante datos ruidosos o redundantes. Por estas razones, se justifica su elección como el modelo final a implementar en el contexto del presente proyecto.

6 Interpretación de Resultados

6.1 Análisis cualitativo de ejemplos

Con el fin de comprender más a fondo cómo el modelo **Random Forest** toma decisiones, se seleccionaron tres ejemplos representativos del conjunto de prueba: dos correctamente clasificados, uno por cada categoría y un caso en el que el modelo cometió un error. Este análisis cualitativo permite observar el comportamiento del modelo ante distintos perfiles de estudiantes.

- **Ejemplo 1: Acierto - Estudiante Avanzado**

  Este caso corresponde a un estudiante que fue correctamente clasificado por el modelo como *Avanzado*. Al revisar sus características, se observa un rendimiento académico consistentemente alto a lo largo de los años. Su promedio general es de 6.9. Sus promedios anuales muestran ser similares durante todo su transcurso académico. Este resultado indica que el modelo es capaz de identificar con alta certeza los perfiles de estudiantes sobresalientes, lo cual es coherente con la lógica pedagógica detrás de la clasificación.

- **Ejemplo 2: Acierto - Estudiante No Avanzado**

  En este segundo ejemplo, el modelo clasificó correctamente a un estudiante como *No Avanzado*. Este estudiante presenta un promedio general inferior a 6.0. Nuevamente, el modelo demuestra un buen desempeño al identificar con precisión los casos con bajo rendimiento histórico. Esto refuerza la idea de que existe una relación directa entre el rendimiento académico acumulado y el nivel asignado al ingresar a 7º básico.

- **Ejemplo 3: Error del modelo**

  El tercer ejemplo corresponde a un error de clasificación. El estudiante fue etiquetado como *No Avanzado*, pero el modelo lo predijo como *Avanzado*. Lo interesante de este caso es que, si bien el promedio general del estudiante es 6.3, el resto de sus indicadores no parecen justificar la clasificación en el nivel superior. Este error sugiere que el modelo podría estar otorgando un peso excesivo al promedio general, lo que provoca que algunos estudiantes con rendimientos intermedios sean clasificados incorrectamente.

6.1.1 Reflexión general

Este análisis cualitativo permite confirmar que el modelo Random Forest opera de manera efectiva en casos extremos, rendimiento muy alto o muy bajo, pero enfrenta dificultades

para categorizar estudiantes con promedios en torno a 6.2–6.4, donde las diferencias no siempre son claras desde un punto de vista cuantitativo.
Además, este tipo de errores puede ser útil desde una perspectiva institucional, ya que permite identificar perfiles fronterizos que podrían beneficiarse de una revisión pedagógica más detallada. En contextos escolares, donde las decisiones de agrupamiento deben considerar múltiples factores, esta información puede complementar el criterio profesional de los docentes.

7 Conclusión

Al explorar los datos del proyecto, se pudo observar un patrón interesante: la separación por niveles al ingresar a 7º básico entre estudiantes *Avanzados* y *No Avanzados* parece tener un efecto positivo en reducir la brecha de rendimiento. Aunque en los años anteriores las diferencias entre ambos grupos eran marcadas, en 7º básico esas distancias tienden a disminuir. Esto sugiere que la categorización por niveles podría estar ayudando a responder mejor a las necesidades de aprendizaje de cada estudiante.
Con ese punto de partida, se desarrolló un modelo de aprendizaje automático para predecir el nivel en que quedará un estudiante al ingresar a 7º básico, a partir de sus antecedentes académicos desde 1º a 6º básico. El proceso incluyó el tratamiento y análisis de los datos, la selección de variables relevantes, y la comparación de distintos algoritmos para encontrar el más adecuado. Finalmente, el modelo *Random Forest* fue escogido por ser considerado preciso y confiable.
Este modelo logró una exactitud cercana al 90 % y mostró un comportamiento consistente tanto con datos de entrenamiento como con datos nuevos, lo que indica que no se limita a "memorizar" sino que realmente aprende patrones útiles. Además, al analizar casos individuales, se vio que el modelo clasifica con seguridad a estudiantes con rendimiento claramente alto o bajo, aunque puede cometer errores cuando los promedios están en un rango intermedio, lo cual es comprensible y aporta información valiosa.
Más allá de los números, este tipo de herramienta puede ser un gran apoyo para docentes y equipos de gestión. Permite anticiparse a situaciones, tomar decisiones informadas y focalizar esfuerzos donde más se necesita. No reemplaza al juicio profesional ni a la mirada humana, pero sí puede potenciarla.
Una idea interesante para continuar este trabajo sería intentar predecir el nivel que alcanzarán los estudiantes en 7º básico, pero utilizando información disponible solo hasta 4º básico. Esto permitiría actuar con mayor anticipación y diseñar estrategias pedagógicas tempranas que acompañen el desarrollo de cada estudiante desde etapas más iniciales, contribuyendo así a cerrar brechas antes de que se amplíen.