DIPLOMATURA
CIENCIA DE DATOS, INTELIGENCIA ARTIFICIAL Y SUS APLICACIONES EN ECONOMÍA Y NEGOCIOS

Análisis descriptivo, predictivo y prospectivo de los resultados de las pruebas PISA para Argentina en 2018: un enfoque de política pública.

Bas Peralta, Benjamín; Tossolini, Lucas

23 de noviembre de 2024

1. Introducción

El presente trabajo se enmarca en la "Diplomatura en Ciencia de Datos: Inteligencia Artificial y sus Aplicaciones en Economía y Negocios", organizada por la Facultad de Ciencias Económicas (FCE) y la Facultad de Matemática, Astronomía, Física y Computación (FAMAF) de la Universidad Nacional de Córdoba (UNC). Por este motivo, se llevó adelante con una mirada evolutiva en cuanto a las técnicas aplicadas, siguiendo el desarrollo de los contenidos del curso.

Hecha esta aclaración, el trabajo, en su faceta aplicada, tiene el propósito de orientar, tomando como medida objetivo el resultado en las pruebas PISA para Argentina en el año 2018, una hipotética asistencia pública a niños de entre 15 y 16 años para intentar mejorar su rendimiento escolar. Con esto en mente, se plantearon dos objetivos complementarios. Primero, uno explicativo que consistió en entender cuáles eran los principales factores influyentes sobre el desempeño de los alumnos en las distintas materias evaluadas por PISA (Matemática, Ciencias y Lengua). En segundo lugar, se planteó un objetivo predictivo para poder determinar cuál será el desempeño de un alumno y poder anticiparse a su situación y actuar a tiempo, en base a los factores antes mencionados.

El trabajo consta de varias etapas. En primer lugar, se realizó una limpieza de los datos y un análisis exploratorio exhaustivo, más focalizado en cada variable por separado. Luego, se continuó con un análisis más profundo, incluyendo metodologías estadísticas multivariadas, de clasificación y de clusterización. Finalmente, se llegó a modelos predictivos complejos.

2. Limpieza de Datos

El conjunto de datos en su totalidad estaba dividida en tres partes. En primer lugar, está la denominada "Base de Estudiantes", la cual contiene los resultados de las pruebas, más ciertos atributos que caracterizan a cada estudiante. Los estudiantes se identifican con `CNTSTUID` y

cuentan con el identificador de la escuela a la que atienden `CNTSCHID` como clave foránea. En segundo lugar, se dispone de la "Base de Escuelas", que contiene atributos de los establecimientos, con su identificador `CNTSCHID`, que permite el cruce con la primera base. Por último, también se dispone de un archivo de codificación, que define los encabezados y las categorías de cada columna de las dos bases anteriores.

La limpieza incluyó en primer lugar un análisis de valores nulos. Con esto, detectamos columnas completamente vacías, las que fueron descartadas. Otras columnas contenían un determinado porcentaje de valores nulos (la base de estudiantes, por ejemplo, contaba con porcentajes de nulos que iban desde 1.1 % hasta 35 %; la de escuelas entre 3 % y 26 %). Al momento de realizar el análisis, se descartaron las columnas con un porcentaje de valores nulos mayor al 10 %. De las columnas restantes, se descartaron las observaciones con al menos un valor nulo en algún atributo. La decisión de eliminar los valores nulos se tomó aún sabiendo que las observaciones con datos faltantes presentaban un leve sesgo si se las analiza por tipo de escuela o nivel socioeconómico del alumno. Otra alternativa podría haber sido la imputación de los valores nulos con alguna técnica, pero no se creyó conveniente dadas las características del problema estudiado.

Además del análisis de datos perdidos, se constató la presencia de valores atípicos. Si bien se detectaron algunas variables con esta situación, se comprobó que dichos valores atípicos no tenían efecto "palanca" sobre el comportamiento general de la variable.

Yendo específicamente a las variables objetivo, las notas finales de cada materia se construyeronn a partir de diez valores incluidos en la base de datos de los estudiantes. Para simplificar el análisis se decidió promediar estos diez valores, construyendo una nota única para Matemáticas (`MATH`), otra para Lengua (`READ`) y otra para Ciencias (`SCIE`).

Finalmente, se trabajó sobre la ponderación de las observaciones. La base de datos de alumnos contenía una variable de ponderación (`W_FSTUWT`). Esta variable hace referencia al peso que se le da a cada alumno de la muestra evaluada para asegurar que los resultados sean representativos de toda la población. En otras palabras, la ponderación simboliza la cantidad de alumnos similares que representa un individuo en la base de datos. Las alternativas que consideramos para trabajar con la ponderación fueron dos. Por un lado, podíamos multiplicar cada fila (que representa un alumno) por su ponderación. Si bien se iban a tener filas duplicadas y el tamaño de la base de datos iba a ser mayor, este método facilitaría el análisis. Por otro lado, podíamos emprender el análisis sin ponderar y considerar las ponderaciones manualmente a posteriori, sin expandir la base. Para este trabajo se optó por la primera alternativa, ya que el tamaño de la base sigue siendo manejable.

Con el preprocesamiento concluido, se procedió al análisis exploratorio.

3. Análisis Exploratorio

El análisis exploratorio se emprendió teniendo en mente el propósito planteado, o sea, intentar identificar variables que influyen en el desempeño de los alumnos. Para ello, se realizaron diversas pruebas estadísticas y gráficas, dividiendo el análisis en dos partes. Por un lado, variables categóricas y, por otro, variables continuas.

Se utilizaron diversas transformaciones de las variables objetivo, dependiendo del análisis realizado. En algunos casos, se emplearon las propias notas promedio en las tres materias (`MATH`, `READ`, `SCIE`). En otras situaciones se construyeron variables dicotómicas que indican la "aprobación" de la materia, para lo cual se recurrió a la documentación oficial de las pruebas PISA ([1]). La escala de notas en cada materia se divide en distintos niveles de competencia, que van del 1 al 6, y PISA suele utilizar como "nivel base de competencia" el nivel 2. Teniendo esto en mente, se definieron los cortes para construir las variables `PASSED_MATH`, `PASSED_READ` y

PASSED_SCIE. Por último, también se recurrió a una variable de desempeño general que indica cuántas materias aprobó el alumno (OVERALL_PERF) bajo los criterios antes mencionados.

El análisis de las variables categóricas incluyó gráficos, test de diferencias de medias y/o test *chi cuadrado*. Se encontraron efectos claros sobre el desempeño de los alumnos según la región en la que vive (SUBNATIO), su estatus migratorio (IMMIG), la repitencia (REPEAT) y el tipo de escuela al que asiste (SCHLTYPE). Por otro lado, el análisis de variables continuas, integrado por diagramas de dispersión y regresiones lineales simples, mostró algún efecto sobre el resultado del nivel educativo (PARED) y laboral de los padres (HISEI), de la educación en la primera infancia (DURECEC), del estatus socioeconómico del niño (ESCS) y de ciertas características de las escuelas. Los gráficos 1 y 2 son ejemplos de estos análisis.

Figura 1: Análisis categórico del Tipo de Escuela (SCHLTYPE) vs. Nota en Matemáticas (MATH).

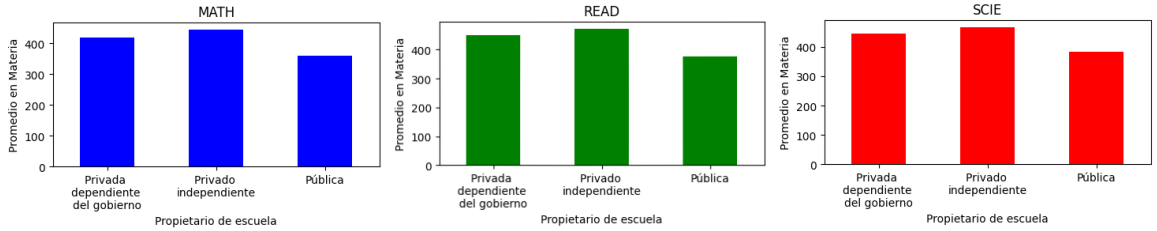
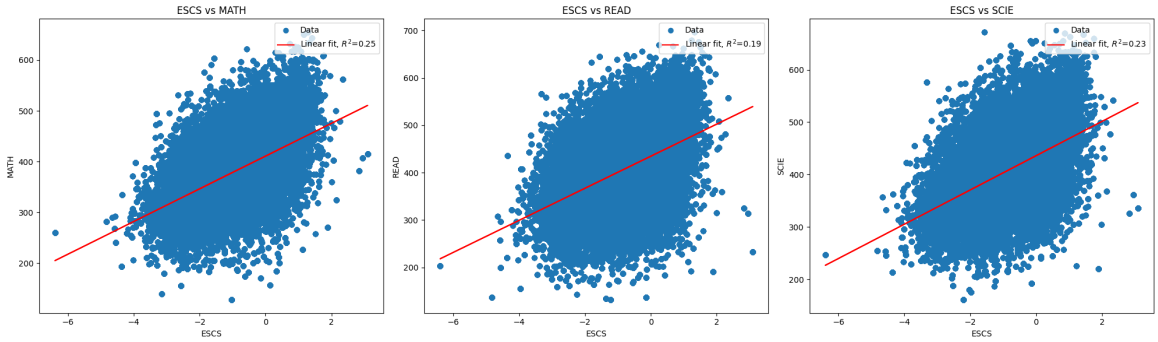


Figura 2: Análisis continuo del Nivel Socioeconómico (ESCS) vs. Nota en Matemáticas (MATH).



4. Análisis Multivariado

4.1. Un Análisis Exploratorio-Descriptivo Avanzado y la primera mirada a la Predicción. La Regresión Multivariada

En esta instancia, y teniendo en cuenta la información obtenida en el análisis anterior, se emplearon regresiones multivariadas para intentar entender el impacto de las distintas variables en conjunto, e ir encaminándonos a un modelo más complejo, en pos de la predicción.

Se probó un modelo MCO (Mínimos Cuadrados Ordinarios) para la nota en Matemática. Dado que los valores de las otras notas (READ, SCIE) están fuertemente correlacionados con MATH, podemos asumir en este primer ejercicio que las regresiones multivariadas darán resultados similares (en otras palabras, utilizaremos el valor numérico de la nota en MATH como *proxy* del resultado general, con el objetivo de lograr eficiencia en el avance del trabajo).

Se incluyeron todas las variables asociadas a los alumnos y las escuelas como explicativas en el modelo (incluso las categóricas convertidas en *dummies*), y se fueron descartando teniendo en cuenta los *test t* de significación individual.

Posteriormente, se realizaron diversas validaciones sobre los resultados de los modelos. En primer lugar, se realizó un análisis de multicolinealidad. Se encontraron correlaciones y, aunque no es necesario, se optó por eliminar algunas variables para simplificar el modelo. Por otro lado, se verificó la especificación lineal del modelo, dando resultados positivos, y se verificaron los supuestos sobre los errores, detectando falta de normalidad. Este problema es atribuible a la presencia de grupos difíciles de detectar con las variables disponibles. Con la intención de probar diversos modelos que colaboren con el entendimiento del problema, se realizó un análisis de endogeneidad. Una vez hecho esto (sin entrar en detalles, ya que están disponibles en el código correspondiente) se optó por utilizar una variable instrumental y realizar una regresión en dos etapas.

El ejercicio realizado con regresiones lineales multivariadas permitió encontrar algunos coeficientes interesantes cuyos valores puntuales confirman algunas intuiciones, pero sobre los que no pueden hacerse más inferencias.

Descubrimos que la nota en Matemática parece responder positivamente a las `CULTPOSS` (posesiones culturales de un alumno en su hogar) y a `ESCS`. Contrariamente a lo intuitivo, responde negativamente a valores altos de `HISEI` y `PARED`. Esto podría deberse a la multicolinealidad detectada, o a un intento del modelo por "compensar" los efectos de las variables. Teniendo este ejemplo en mente, se observó que los resultados obtenidos de manera individual para cada variable, no necesariamente se reflejan en el ejercicio multivariado. Por este motivo, en la próxima sección se avanzará considerando las variables `CULTPOSS`, `ESCS`,`HISEI` y `PARED` como un agregado denominado "situación en el hogar" del alumno. En base a esta característica, se intentarán encontrar grupos de estudiantes similares, y analizar su desempeño.

Siguiendo con los resultados multivariados, la nota en Matemática también parece responder positivamente a `TEACHBEHA` (comportamiento del profesor), `DISCLIMA` (clima disciplinario en el aula) y `SCHSIZE` (tamaño de la escuela medido en número total de alumnos). De manera similar a lo mencionado para la "situación en el hogar", se intentará profundizar más en la "situación en la escuela" (las mencionadas variables `TEACHBEHA`, `DISCLIMA` y `SCHSIZE`), con el objetivo de comprender mejor las implicancias sobre el resultado de los niños.

Por último, las regresiones multivariadas mostraron que los coeficientes de las variables *dummy* confirman los comportamientos observados en la parte exploratoria.

Es necesario realizar aclaraciones sobre el uso de ciertas metodologías que no arrojaron resultados útiles. En este sentido, se intentó avanzar en un Análisis de Componentes Principales, pero se llegó a la conclusión de que sus resultados no resultan apropiados dados los objetivos del trabajo.

4.2. Profundizando en la Predicción: Clusterización

Continuando en la línea de profundizar la caracterización y descripción del resultado de los alumnos, y teniendo en cuenta que el modelo de regresión multivariada no nos aportó suficiente información en pos de nuestro objetivo predictivo, se consideró que la base de datos PISA se prestaba para llevar a cabo un análisis de *clustering*.

Este análisis permitió encontrar diferentes agrupaciones naturales de las observaciones que sirvieron de guía para las primeras recomendaciones. La agrupación se hizo teniendo en cuenta las variables asociadas con la "situación en el hogar": `ESCS`, `HISEI`, `CULTPOSS`, `PARED`. A partir de ellas, se generaron cinco clusters de alumnos, que demostraron tener comportamientos diferentes en los resultados académicos.

El gráfico 3 muestra que, como era esperado, los dos clusters extremos en cuanto a la "situación en el hogar" tienen resultados casi opuestos en las calificaciones. En cambio, los clusters construidos según la "situación en la escuela" (`SCHSIZE`, `DISCLIMA` y `TEACHBEHA`) no muestran una regularidad en el espectro de notas, por el contrario parecen distribuidos casi de

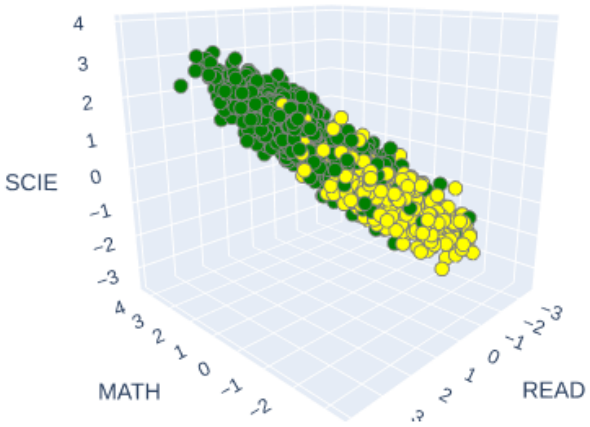


Figura 3: Ejemplo de la diferencia de los clusters amarillo (peores indicadores de "situación en el hogar") y verde (mejores indicadores).

manera aleatoria.

Dicho esto, en los términos de la elección pública, y la restricción presupuestaria, parece más apropiado recomendar una política de intervención en los hogares. Vale la pena recordar que también existen otras agrupaciones predefinidas en la base (por ejemplo, escuela pública o privada, repitencia, región geográfica) que demostraron tener efectos significativos en los análisis. Estas características también deberían ser tenidas en cuenta a la hora de priorizar la asistencia.

5. Predicción: Clasificación Supervisada

La siguiente pregunta a responder es *¿cómo se puede predecir a qué alumno canalizar ayuda para mejorar su situación en el hogar?*. Para ello, se definió un criterio que sirva a los fines de la clasificación. Este fue: "el alumno supera (o no) el valor que PISA considera como el nivel básico de competencia en cada materia", es decir el alumno "aprobará" o no.

Para llegar a la predicción final, se probaron varios modelos. En la elección del mejor de ellos se tuvo en cuenta una métrica que reflejara el interés primordial de este trabajo: identificar a los alumnos que potencialmente desaprobarían. Por ello, se intenta minimizar la cantidad de falsos positivos, es decir la cantidad de casos en las que el modelo predice que el alumno va a aprobar cuando en realidad no aprueba, y maximizar la cantidad de verdaderos negativos. Dicho esto, la métrica a la que prestamos más atención es la de "*Recall*" para valores negativos, también llamada "*True Negative Rate*" (TNR), aunque también tendremos en cuenta la "*Accuracy*", o sea la precisión global. Ambas métricas están dadas por las siguientes fórmulas:

$$\text{Precisión Global} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}, \quad \text{TNR} = \frac{\text{VN}}{\text{VN} + \text{FP}},$$

donde VP representa los verdaderos positivos, VN los verdaderos negativos, FP los falsos positivos y FN los falsos negativos.

Al obtener los conjuntos de entrenamiento y de testeo, notamos que ambos estaban desbalanceados: 65 % con clase 0 (desaprobación) y 35 % con clase 1 (aprobación). Por esta razón, decidimos balancear los datos. Esto significa, igualar la cantidad de observaciones con presencia

de la situación deseada (aprobó) y con ausencia de ella (no aprobó). En el caso presentado, la segunda categoría es más abundante y, además, es la más interesante a detectar a los fines de este trabajo: aquellos niños que, dadas sus características, no aprobarían las materias evaluadas. Para realizar el balanceo, existen varias maneras: sobremuestreo de clase minoritaria, submuestreo de clase mayoritaria, o aplicar pesos a las clases a la hora de hacer el entrenamiento. Nosotros empleamos la última, dado que era una funcionalidad ya incorporada en la librería utilizada. Consiste en, a la hora de entrenar el modelo, darle un mayor peso a errores cometidos en la clase minoritaria que a errores en la clase mayoritaria. Estos pesos son inversamente proporcionales a la cantidad de instancias que existen de una clase determinada.

Otra transformación que aplicamos sobre los conjuntos fue la estandarización de las variables continuas. Para llevarla a cabo, tomamos la media y desviación estándar de cada variable de los datos de entrenamiento, y con ellas aplicamos la normalización sobre el conjunto de entrenamiento y de testeo. Hacer esto evita el problema de "fuga de datos", en donde el conjunto de entrenamiento se ve afectado por información del de testeo.

A continuación, describimos los diferentes modelos utilizados y las métricas obtenidas, haciendo hincapié en la precisión global y el TNR.

5.1. Clasificación con Discriminante lineal

Luego de realizar varias pruebas, se llegó a la conclusión de que el modelo lineal, incluyendo todas las variables, era el que mejores resultados arrojaba. La precisión global fue del 78 %, y el TNR fue del 87 %.

Si bien aquí no llevamos a cabo balanceo ya que la librería no lo soportaba, y obtuvimos peores métricas en general para la clase minoritaria, nos valimos de la simplicidad de este modelo para analizar un poco los pesos asignados por la recta a cada variable. De aquí confirmamos lo que ya veníamos observando: el tipo de escuela, el nivel socioeconómico de los padres, y la repitencia son los factores con mayor peso.

5.2. Clasificación con Discriminante logístico

Los resultados obtenidos con el discriminante logístico son bastante similares a los del discriminante lineal. La precisión global fue del 75 % y el TNR del 74 %. Son un poco más bajas que en el modelo anterior, pero a la vez más reales ya que aquí sí pudimos aplicar balanceo.

La última etapa del análisis predictivo incluye árboles de decisión y una red neuronal con perceptrón multicapa, para intentar mejorar la predicción obtenida con el discriminante logístico.

5.3. K vecinos más cercanos

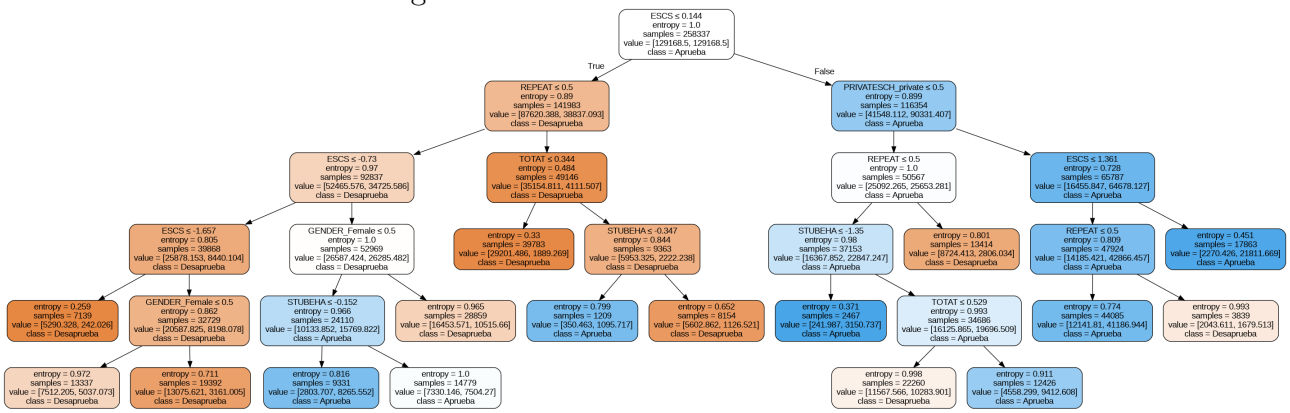
Probando diferentes valores para la cantidad de vecinos en la cual fijarse al hacer la predicción, descubrimos sorprendentemente que tomar 3 vecinos da valores muy altos de todas las métricas en general. Sin embargo, la gran desventaja de esta técnica es su alto costo computacional a la hora de hacer la predicción.

5.4. Árbol de decisión

Podría decirse que la idea detrás de entrenar un árbol de decisiones es el aprendizaje de los condicionales que llevan a una respuesta: ¿el alumno aprobó o no?. Esto, aunque sirve a los fines predictivos, quizás resulta más útil para el objetivo orientado a entender los *drivers* del desempeño de los estudiantes, dada la versatilidad gráfica de la técnica (ver gráfico 4).

Para la creación del árbol, y llevando a cabo mediciones de la precisión global obtenida y de la TNR variando los hiperparámetros de "profundidad máxima" y "máxima cantidad de nodos", se optó por estimar un árbol de profundidad 16. Fijada la profundidad, el mejor resultado fue obtenido con un número máximo de hojas de 16. Ante estas restricciones, el árbol entrenado terminó contando con una profundidad de 5 y una cantidad de hojas de 16. La métrica TNR del árbol obtenido es de 77 % y la precisión global del 75 %.

Figura 4: Árbol de decisión resultante.



Se confirmaron la mayoría de los comportamientos observados en los análisis anteriores. Para los alumnos que tienen un peor `ESCS` (lado izquierdo del árbol), aparece la repitencia como un factor decisivo, mientras que para los que tienen un mejor `ESCS`, el tipo de escuela aparece como el principal determinante. Es interesante notar que la variable `STUBEHA`, que representa el comportamiento del alumno, también aparece como una característica importante, lo cual no lo habíamos identificado hasta el momento.

5.5. Redes Neuronales: Perceptrón Multicapa de Clasificación

Finalmente, se llegó al modelo predictivo definitivo: altas métricas y costo computacional efectivo. Se entrenó un perceptrón multicapa con tres capas ocultas, las dos primeras de 30 neuronas y la tercera de 10. La función de activación de las capas fue ReLU, la función de pérdida fue *log-loss*, y el algoritmo de optimización fue Adam, con un número máximo de épocas de entrenamiento de 200. El tiempo de entrenamiento fue de aproximadamente 10 minutos. Con este modelo, obtuvimos una increíble precisión en las predicciones, y además el TNR fue de 99 %.

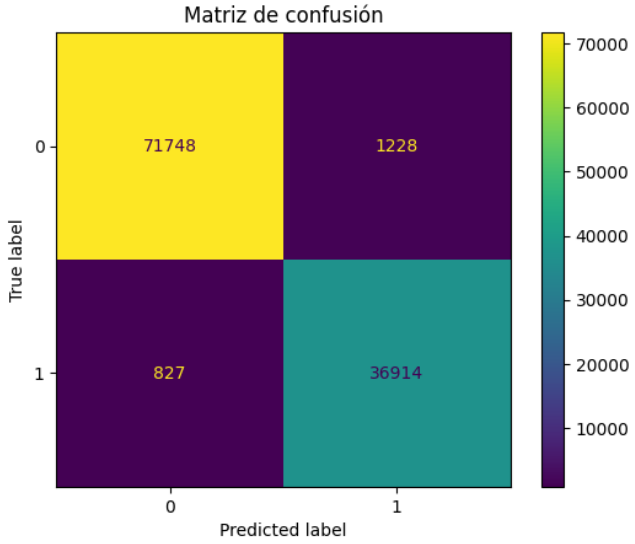
5.6. Resumen de Modelos

6. Conclusiones

Las primeras instancias de la investigación, tales como el análisis exploratorio, el análisis multivariado, y la clusterización nos permitieron lograr nuestro objetivo explicativo de intentar determinar cuáles son los factores más influyentes en el rendimiento escolar de un alumno.

En primer lugar, ya en el análisis exploratorio se encontraron relaciones relevantes. Por ejemplo, los alumnos de Tucumán, inmigrantes y aquellos que repitieron obtienen peores calificaciones, mientras que los de escuelas privadas, mejores. Por otro lado, los alumnos con un nivel socioeconómico más alto (representado principalmente por la variable `ESCS`, y en menor

Figura 5: Matriz de confusión del Perceptrón Multicapa



Modelos	Recall para valores 0 (TNR)	Precisión global
Discriminante Lineal	73 %	74 %
Discriminante Logístico	74 %	75 %
K vecinos más cercanos	**99 %**	**99 %**
Arbol de decisión	77 %	75 %
Perceptrón Simple	70 %	66 %
Perceptrón Multicapa	**98 %**	**98 %**

Cuadro 1: Comparación de modelos basada en TNR y precisión global.

medida por otras como `PARED` y `HISEI`) tienen un mejor desempeño; lo mismo que aquellos con un clima educativo más propicio (explicado por variables como `TEACHBEHA`, `DISCLIMA` y `SCHSIZE`).

Cuando se avanzó un poco más en el análisis, se encontraron clusters socioeconómicos claros, con una evidente contraposición en los resultados obtenidos. Se confirmó que tanto "la situación en el hogar" como "en la escuela" tienen un impacto en el resultado. Sin embargo, a los fines de la política pública (foco de este análisis), resulta más conveniente atacar lo más evidente: la "situación en el hogar".

Al pasar ya a las etapas finales de la investigación, las técnicas utilizadas, que ya no brindan tanto poder explicativo como las empleadas anteriormente, contribuyeron a alcanzar nuestro segundo objetivo. De este modo, procedimos a la predicción del desempeño de los estudiantes, lo que podría guiar con mayor eficiencia la aplicación de la política mencionada. En este sentido, el modelo con mejores resultados fue el Perceptrón Multicapa.

Como apreciación final, consideramos que el método evolutivo y fuertemente analítico de este trabajo permitió llegar a un modelo predictivo con alta precisión, sin perder potencia explicativa (la que hubiera estado ausente si se hubiera recurrido de inmediato a modelos "oscuros", como una red neuronal). Si se hubiera otorgado un peso excesivo a los modelos predictivos presentados en la última sección, se habría comprometido la comprensión del fenómeno y la decisión de abordar el problema desde el entorno familiar y hogareño. No obstante, estos modelos son útiles para identificar a los estudiantes con mayor probabilidad de desaprobar.

Referencias

[1] OECD: PISA 2009 Results: What Students Know and Can Do (2010). https://doi.org/https://doi.org/https://doi.org/10.1787/9789264091450-en, `https://www.oecd-ilibrary.org/content/publication/9789264091450-en`, en 2018 en Argentina las pruebas PISA se hicieron con papel y lápiz, por lo que debimos consultar este documento, y no el de 2018.