Artículo Original

CatBoost: Aprendizaje automático de conjunto para la analítica de los factores socioeconómicos que inciden en el rendimiento escolar

CatBoost: Ensemble machine learning for the analysis of socioeconomic factors that affect school performance

Jorge Iván Pincay-Ponce[1,2], Armando E. De Giusti[2], Diana Alexandra Sánchez-Andrade[3], Juan Alberto Figueroa-Suárez[1]

[1] Universidad Laica Eloy Alfaro de Manabí, Manta, Ecuador
[2] Universidad Nacional de La Plata, Facultad de Informática, La Plata, Argentina
[3] Universidad de Guayaquil, Guayaquil, Ecuador

jorge.pincay@uleam.edu.ec, degiusti@lidi.info.unlp.edu.ar, diana.sancheza@ug.edu.ec, juan.figueroa@uleam.edu.ec

Recibido: 09/04/2023 | **Aceptado**: 08/08/2023

Cita sugerida: J. I. Pincay-Ponce, A. E. De Giusti, D. A. Sánchez-Andrade, J. A. Figueroa-Suárez, "CatBoost: Aprendizaje automático de conjunto para la analítica de los factores socioeconómicos que inciden en el rendimiento escolar," *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, no. 38, pp. 31-39, 2024. doi:10.24215/18509959.38.e3.

Resumen

El rendimiento académico de los niños es una importante tarea para las escuelas y es de atracción desde el campo de la ciencia de datos que atiende esta problemática multifactorial con diversas técnicas de minería de datos sobre conjuntos de datos cada vez más completos que abordan factores socioeconómicos como posibles condicionantes. Presentamos un método que mejora la Exactitud de la predicción del rendimiento escolar combinando la aplicación del modelo de aprendizaje automático en conjunto CatBoost con la explicación y mejora de la transparencia de la clasificación que efectúa, mediante la puntuación de las características con base en los valores SHAP (SHapley Additive exPlanations). Se dispone de cuatro tipos de promedios: Domina los aprendizajes requeridos (DAR), Alcanza los aprendizajes requeridos (AAR), Próximo a alcanzar los aprendizajes requeridos (PAAR) y No alcanza los aprendizajes requeridos (NAAR). Cómo los tipos de promedios PAAR y NAAR constituyen clases minoritarias fueron balanceados respecto de las clases mayoritarias DAR y AAR. Se alcanzó una Exactitud y Precisión del 91%. Las características de mayor impacto en la predicción son las habilidades sociales, la ocupación del padre, ingreso familiar, género, posible discapacidad, comportamiento, estructura familiar, número de hermanos, entre otros.

Palabras clave: Aprendizaje automático; CatBoost; Shapley; Rendimiento académico; Métodos de ensamble.

Abstract

The academic performance of children is an important task for schools and is attractive from the field of data science that addresses this multifactorial problem with various data mining techniques on increasingly complete data sets that address socioeconomic factors such as possible conditions. We present a method that improves the accuracy of the prediction of school performance by combining the application of the ensemble learning algorithms CatBoost with the explanation and improvement of the transparency of the classification that it performs, by scoring the characteristics based on the SHAP values. (Shapley Additive exPlanations). Four types of averages are available: Master the Learning Requirement (DAR), Meet the Learning Requirement (AAR), Close to Meeting the Learning Requirement (PAAR), and Not Meet the Learning Requirement (NAAR). How the PAAR and NAAR types of averages constitute minority classes were balanced with

respect to the DAR and AAR majority classes. An accuracy and precision of 91% were achieved. The characteristics with the greatest impact on the prediction are social skills, the father's occupation, family income, gender, possible disability, behavior, family structure, and number of siblings, among others.

Keywords: Machine learning; CatBoost; Shapley; Academic performance; Assembly methods.

1. Introducción

La obtención masiva de datos está configurando diferentes sectores debido al creciente número de sistemas informáticos que almacenan datos de diferentes fuentes. El sector de la educación es uno de los más involucrados, dado el potencial subyacente en los datos para apoyar la enseñanza y el aprendizaje efectivo que mejoren las formas en que las generaciones futuras construirán la realidad con y a través de los datos [1], favoreciendo a la calidad, la eficiencia y la inclusión educativa escolar, sin embargo, en muchas escuelas se recopila cantidades considerables de datos porque las personas los han estado almacenando durante muchos años, en lugar de por unas razones intencionales de analítica de relaciones entre los puntajes y factores que los afectan de modo clave [2], [3], [4].

Para muchos niños, dejar la escuela es el paso final en un largo proceso de desconexión gradual y reducción de la participación en el currículo escolar formal, así como en la vida social de la escuela. Como el detonante de la deserción puede asociarse a la reprobación o no aprobación de una o más materias a causa de un insuficiente rendimiento cuantitativo o cualitativo, en la actualidad las escuelas toman como referencia una serie de políticas y programas para evitar el fracaso escolar, sin embargo, este es un problema multifactorial, factores entre los que están los socioeconómicos [5].

La aplicación de técnicas de aprendizaje automático para predecir el rendimiento de los estudiantes, con base en sus antecedentes académicos y factores socioeconómicos ha demostrado ser una herramienta útil para prever buenos y malos resultados en los distintos niveles educativos, ejecutar medidas tempranas para mejorar el resultado del aprendizaje y orientar los cambios en las políticas académicas [6]. Muchos de los datos socioeconómicos referidos son de naturaleza categórica o de texto.

Una de las maneras de abordar la problemática con aprendizaje automático es mediante el aprendizaje en conjunto o de algoritmos ensamblados, en este método se utilizan varios algoritmos al mismo tiempo, para lograr un rendimiento predictivo más alto que si se utilizara un algoritmo individual por sí mismo, aunque aumente la complejidad computacional en comparación con los clasificadores individuales. Estos algoritmos construyen muchos árboles en el proceso y realizan la predicción final con base en todos los árboles [7]. Ciertamente los hay de varios tipos: Voting, Bagging y Boosting.

En esta investigación se emplea CatBoost que se corresponde con Boosting, traducido como impulso del gradiente. En Boosting los modelos aprenden secuencialmente con las primeras instancias que ajustan modelos simples a los datos y luego analizan los datos en busca de errores. En los árboles consecutivos el objetivo es mejorar la precisión del árbol anterior, porque cuando una instancia se clasificó erróneamente ante una hipótesis, su peso aumenta para que la siguiente hipótesis tenga más probabilidad de clasificación correcta. En los algoritmos de Boosting, los modelos simples son utilizados secuencialmente, es decir, cada modelo simple va delante o detrás de otro modelo simple [7], [8].

El objetivo de este estudio es contribuir a mejorar la utilidad de las métricas de rendimiento mediante el algoritmo de aprendizaje en conjunto CatBoost y el balanceo ponderado de datos, a la hora de predecir el rendimiento de los niños estudiantes de dos escuelas ecuatorianas.

En las siguientes subsecciones se presentan detalles fundamentales de CatBoost, SHAP y Smote.

1.1. Catboost

A menudo, los algoritmos de aprendizaje automático son complejos porque requieren de largos tiempos de CPU para ser entrenados y muchas veces carecen de herramientas que puedan ayudar a explicar los valores de las características que respaldan sus decisiones, así como el entrenamiento del modelo en sí, entre otras razones [8]. Buena parte de ese tiempo sucede durante el tratamiento de características categóricas que requieren de ser transformadas como numéricas con, por ejemplo, una codificación en caliente, que conlleva generar más columnas si es que se dispone de muchas columnas en un inicio y de valores distintos en cada una de ellas [9].

El método resulta útil especialmente cuando solo hay dos posibilidades que se valorarán en 0 o 1, pero cuando el dominio o posibilidades de los valores es extenso los modelos tardarán demasiado en entrenarse y hasta de comprenderse por parte del usuario, especialmente si se ejecutan con frecuencia en un entorno de producción. El procesamiento de características categóricas permite un entrenamiento más rápido al tiempo de favorecer la interpretabilidad y transparencia de su decisión ante los usuarios interesados [10].

El algoritmo CatBoost se clasifica dentro de los modelos de Boosting, es de código abierto, surgió en 2017. Proporciona un aumento de gradiente que intenta resolver las características categóricas utilizando una alternativa impulsada e imparcial por permutaciones para construir las divisiones de árboles CART(Classification And Regression Tree) y elegir los valores hojas a partir de conjuntos de datos diferentes en cada una de las iteraciones [8]. También admite objetivos binomiales y continuos. Sus parámetros se ajustan automáticamente contribuyendo a su puesta en producción en menor tiempo posible [11].

CatBoost utiliza la técnica de codificación de destino ordenada, es decir, si los datos no tienen un registro de tiempo el algoritmo crea aleatoriamente un tiempo artificial para cada punto de datos [8].

1.2. Interpretabilidad

Para facilitar la interpretación del modelo, se utilizaron valores SHAP, acrónimo de Shapley Additive Explanations. Los valores SHAP suplen la problemática de que los modelos tienen un valor de salida que no es interpretable con facilidad porque se centran en el "¿cuánto?" del problema, es decir, que se desconocen las razones de sus salidas. Los valores SHAP ilustran cómo afecta cada característica a la predicción, aún sobre lo complejo que son métodos como el refuerzo por gradiente (tal es el caso de CatBoost) o las redes neuronales. La interpretación de modelos de aprendizaje automático se ha beneficiado del marco SHAP propuesto en 2016 por Scott Lundberg y Su-ln Lee [12], porque SAHP resultó una forma fácil y teóricamente sólida de entender las predicciones de cualquier modelo.

1.3. Datos desbalanceados

En general, el objetivo de un algoritmo de aprendizaje automático en una tarea de clasificación es producir un clasificador que maximice la Exactitud, entendida como el número total de instancias clasificadas correctamente. Sin embargo, esto no es suficiente para producir clasificaciones ideales ante el problema de disponer de un conjunto de datos desequilibrados, porque la Exactitud por sí sola conduciría a conclusiones incorrectas puesto que en ella se considera a la Exactitud general y no a la de cada clase [13], [14]. La literatura existente sugiere varias técnicas para sobrellevar tales casos, en aras de no perder instancias para el análisis como sucede con el submuestreo de clases mayoritarias [13], se optó por el empleo de SMOTE para el sobremuestreo de las indicadas clases minoritarias.

SMOTE significa Synthetic Minority Oversampling Technique o Técnica de sobremuestreo de minorías sintéticas, es una técnica de aprendizaje automático que realiza el aumento de datos mediante la creación de puntos de datos sintéticos ligeramente diferentes de los puntos de datos originales, es decir, sin generar duplicados, con ello, se evita que el modelo casi nunca prediga clases minoritarias. Las instancias se definen típicamente utilizando los k vecinos más cercanos a una instancia particular de las clases minoritarias [15], [16].

2. Metodología

En esta sección, demostramos el enfoque para utilizar el método de aprendizaje en conjunto CatBoost, considerando el problema del desbalanceo de clases, así como la importancia de transparentar las decisiones de los modelos empleando los valores SHAP.

Respecto de los pasos seguidos se tomó como referencia los ciclos de vida más populares para estos casos: 1) Preparación de los datos, (2) Construcción del modelo y (3) Evaluación del modelo.

El análisis se realizó con el software Orange 3.34, adicionalmente se empleó la librería imbalanced-learn que se basa en scikit-learn [14]. Respecto del computador se empleó uno con procesador AMD Ryzen 5 de 2.10 GHz, 32 GB de RAM, 4 cores, 8 procesadores lógicos, 4 MB de memoria caché L3 y Disco SSD de 1 TB.

2.1. Descripción del conjunto de datos

Los datos objeto de estudio corresponden a una muestra transversal del periodo lectivo 2019 de dos escuelas de Ecuador, dado que la cantidad de datos entre una y otra difería considerablemente, no se construyó un análisis comparativo, sino uno con base en la información consolidada que totaliza 6808 instancias de calificaciones y 88 columnas que lo describen[1]. El análisis giró en torno a cada registro de calificaciones y no de cada alumno, porque en el sistema escolar ecuatoriano las bajas calificaciones en una materia, simplificadas como rendimiento académico, pueden llegar a determinar la reprobación del año básico cursado por el alumno [17]. Para el análisis se excluyó códigos de identificación y nombres de los alumnos.

En relación con el análisis de la incidencia de los factores socioeconómicos son de especial relevancia los siguientes datos: Estado civil del padre, madre y del representante; escolaridad del padre, madre y del representante; ocupación del padre, madre, y del representante; parentesco del representante, número de hermanos, estructura familiar, ingreso monetario del hogar, servicios de energía eléctrica, agua potable regularizado, alcantarillado, internet, televisión por cable y telefonía celular regularizados; computador en casa, discapacidad, año de ingreso a la escuela, materia dificultosa, procedencia o no desde otra institución, repetidor o no de año básico y enfermedad auto reportada. La **Tabla 1** muestra la correspondencia de las calificaciones con las cuatro clases [18].

Tabla 1. Escala de calificaciones propuesta por el Ministerio de Educación de Ecuador

Nº	Siglas	Significado	Rango numérico
1	**DAR**	Domina los aprendizajes requeridos	desde 9.00 hasta 10.00
2	**AAR**	Alcanza los aprendizajes requeridos	desde 7.00 a 8.99.
3	**PAAR**	Próximo a alcanzar los aprendizajes requeridos	desde 4.001 a 6.99.
4	**NAAR**	No alcanza los aprendizajes requeridos	menos o igual a 4.

2.2. Preparación de los datos

La preparación de datos es un paso en el que los datos recopilados se convierten a un formato tabular adecuado. A nivel de filas se consolidó las filas de los datos de las dos escuelas, se removió duplicados, se removió valores

atípicos, se sobremuestreó las clases minoritarias. A nivel de valores se imputó y cálculos de valores faltantes, se cambió ciertos valores y se escaló los datos pertinentes. A nivel de columnas se excluyó características redundantes y se generó otras 26 columnas que forman parte de las 88 mencionadas en la sección que precede[2].

Tras este paso, se efectuó el sobremuestreo ponderado con SMOTE, que incrementó los datos en aproximadamente un 37% tal como se observa en la **Tabla 2**. La razón de utilizar la técnica de sobremuestreo es la estructura sesgada de los métodos de árbol de decisión CART en los que se basa CatBoost, mismos que tienden a efectuar clasificaciones más orientadas hacia las clases mayoritarias en los conjuntos de datos desbalanceados, debido a ello, se producen imprecisiones en la predicción de la clase mayoritaria; como resultado, la clase minoritaria se clasifica erróneamente en comparación con la clase mayoritaria. Para SMOTE se siguió estos pasos: 1) Se submuestreó o redujo la cantidad de filas de la clase mayoritaria AAR pasando de 3943 instancias a 3076 y (2) Se sobremuestreó a las clases minoritarias PAAR y NAR.

Tabla 2. Instancias entrantes y salientes por clases en SMOTE

	DAR	AAR	PAAR	NAAR	Total
Entrada	2552	3076	227	86	5941
Salida	2552	3076	1500	1000	8128

2.3. Construcción del modelo

Tras varias pruebas de optimización de hiperparámetros la configuración se fijó como se muestra en la **Tabla 3**.

Tabla 3. Personalización de hiperparámetros para CatBoost

Parámetro	Valor
Número de árboles (CART)	100
Tasa de aprendizaje	0.50
Función de regularización	Lambda 1
Límite de profundidad de cada árbol	10
Fracción de características para cada árbol	100%
Entrenamiento replicable	Sí

La **Tabla 3** muestra el número de estimadores débiles CART que se ha empleado. La tasa de aprendizaje establece que tanto se actualizan los pesos en cada iteración en un rango de 0 a 1, un valor cercano a 1 podría cometer errores y por ende un modelo de predicción inadecuado, pero un valor cercano a 0 podría tardar el entrenamiento de CatBoost, por tanto, se fijó un término medio.

Respecto de Lambda 1, L1, también se conoce como desviaciones mínimas absolutas, errores mínimos absolutos. Básicamente se trata de minimizar la suma de las diferencias absolutas $\boldsymbol{S}$ entre el valor objetivo y_i y los valores estimados $f(x_i)$: $S = \sum_{i=1}^{n} |y_i - f(x_i)|$. El principal beneficio de esta regularización es mitigar el sobreajuste controlando el proceso de aprendizaje de CatBoost agregando otro término a la función de pérdida (costo) que se busca minimizar. L1 se define como $\alpha \sum(valores\ al\ cuadrado\ de\ los\ coeficientes)$, Alfa es un hiperparámetro implícito en el caso de Orange, que controla la fuerza de regularización, debe ser un float positivo. El valor predeterminado es 1. Los valores más grandes de alfa implican una regularización más fuerte, es decir menos sobreajuste, los más pequeños implican una regularización débil, es decir, sobreajuste [19]. Cuando se dispone de conjuntos de datos pequeños, el gradiente tiende a sobre ajustarse rápidamente.

2.4. Evaluación del modelo

El muestreo empleado en el modelo es aleatorio estratificado con el 80% de instancias para el entrenamiento en cada uno de los 10 estratos que se fijó. Se prescindió de la validación con un conjunto de datos de prueba porque los datos para el entrenamiento son sobremuestreados y los de prueba no, entonces era poco probable que se tuviese una proporción representativa de instancias con clases minoritarias al momento de probar y las que hubiese tenderían a reportar altos valores de Exactitud, pero poco realistas.

Luego, la métrica más utilizada en un problema de clasificación como el presente es la Exactitud. Para ejemplificar TP significa verdadero positivo, que se refiere al conjunto de instancias para las que la predicción del modelo era correcta. El negativo verdadero (NT) es el conjunto de instancias cuya predicción fue correcta; el falso positivo (FP) es el número de instancias cuya predicción fue incorrecta y el falso negativo (FN) es el número de instancias cuya predicción fue incorrecta.

Por tanto, la Exactitud se calcula como $Exactitud = \frac{TP_1+TP_2+\cdots+TP_n}{TP+FP}$ [20].

La Precisión es una medida de corrección que indica cuántas predicciones son realmente positivas de todo el total positivo predicho. Se calcula como $Precisión = \frac{TP}{TP+FP}$ [20].

El Recuerdo es una medida de observaciones reales que se predicen correctamente, también se conoce como sensibilidad, se calcula como $Recuerdo = \frac{TP}{TP+FN}$ [20].

Al disponer de sobremuestreo de clases minoritarias se espera aumentar el valor de esta métrica entendida también como capacidad de generalización del modelo para con nuevos datos.

3. Resultados

Con base en las configuraciones precedentes se obtuvieron los resultados globales de la clasificación presentados en la **Tabla 4**.

Tabla 4. Exactitud, precisión, recuerdo y tiempos obtenidos. El punto se emplea como separador decimal

Métrica	Con SMOTE	Sin SMOTE
Exactitud	0.915	0.890
Precisión	0.916	0.890
Recuerdo	0.915	0.890
Tiempo de entrenamiento ms	452.550	281.928
Tiempo de prueba ms	2.580	2.082
Instancias disponibles	6808	16260

La matriz de confusión, ver **Tabla 5**, muestra los resultados de la Exactitud para cada clase. En ella se aprecian situaciones como que 9.9% de instancias de la clase Alcanza los aprendizajes requeridos (AAR) se clasifican erróneamente como de la clase que Domina los aprendizajes requeridos (AAR), en tanto que un 9.5% de las clases DAR se clasifican como AAR. Los niveles más altos de CatBoost ocurren en las clases que representan problemas de rendimiento en sí: PAAR y NAAR. En general, los errores de clasificación suelen ocurrir con la clase que representa al promedio inmediato superior o inmediato inferior.

Tabla 5. Matriz de confusión con datos sobremuestreados

	DAR	AAR	PAAR	NAAR	Total
DAR	90.3%	9.9%	0.4%	0.2%	**5174**
AAR	9.5%	88.1%	3.4%	0.1%	**6165**
PAAR	0.2%	1.8%	96.2%	0.5%	**2935**
NAAR	0.1%	0.2%	0.0%	99.2%	**1986**
Total	**5021**	**6341**	**2914**	**1984**	**16260**

En la **Figura 1** se han ordenado de modo descendente a las características de acuerdo con como contribuyen al poder predictivo de CatBoost. La asignatura, los proyectos escolares que el caso de este estudio evalúan las habilidades sociales [21], la ocupación del padre que se relaciona con el ingreso familiar, el género del alumno, tener discapacidad o no, el comportamiento, la estructura familiar, número de hermanos, haber reprobado cursos o no, entro otros, son las características de más poder en la predicción global. Nótese que, aunque los enfoques de conjunto, como CatBoost, pueden alcanzar altos índices de Exactitud y Precisión, son difíciles de entender, por lo que estos modelos suelen verse como cajas negras, no obstante, al incluirse los valores SHAP para calcular los valores de significancia de las características se consigue mejorar la interpretabilidad del modelo.

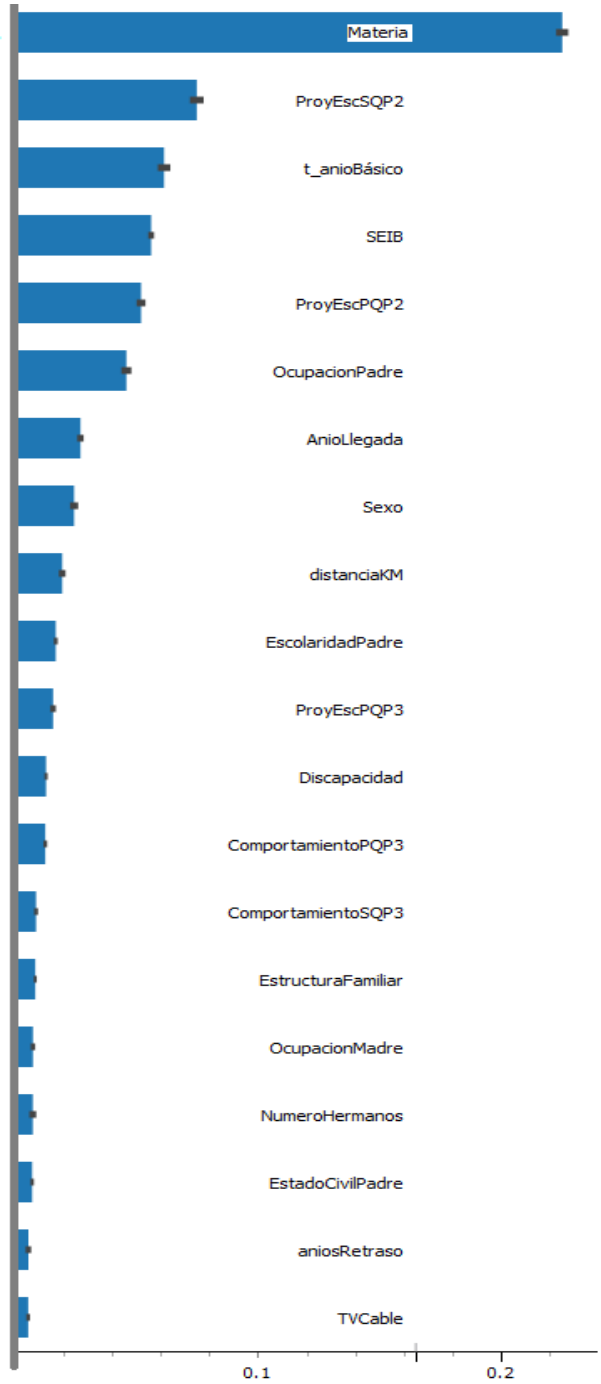

Figura 1. Importancia de las características en la Exactitud de la clasificación lograda

En la **Figura 2** se han ordenado de modo descendente a las características de acuerdo con como contribuyen al poder predictivo de CatBoost respecto de la clase Domina los aprendizajes requeridos, DAR. Este es el tipo de promedio más alto posible. Para la clasificación de los promedios de esta clase impactan más en la decisión: materia, proyectos escolares, año básico (que indirectamente representa la edad del niño), ocupación del padre (que indirectamente representa el ingreso familiar), comportamiento, escolaridad del padre, entre otros.

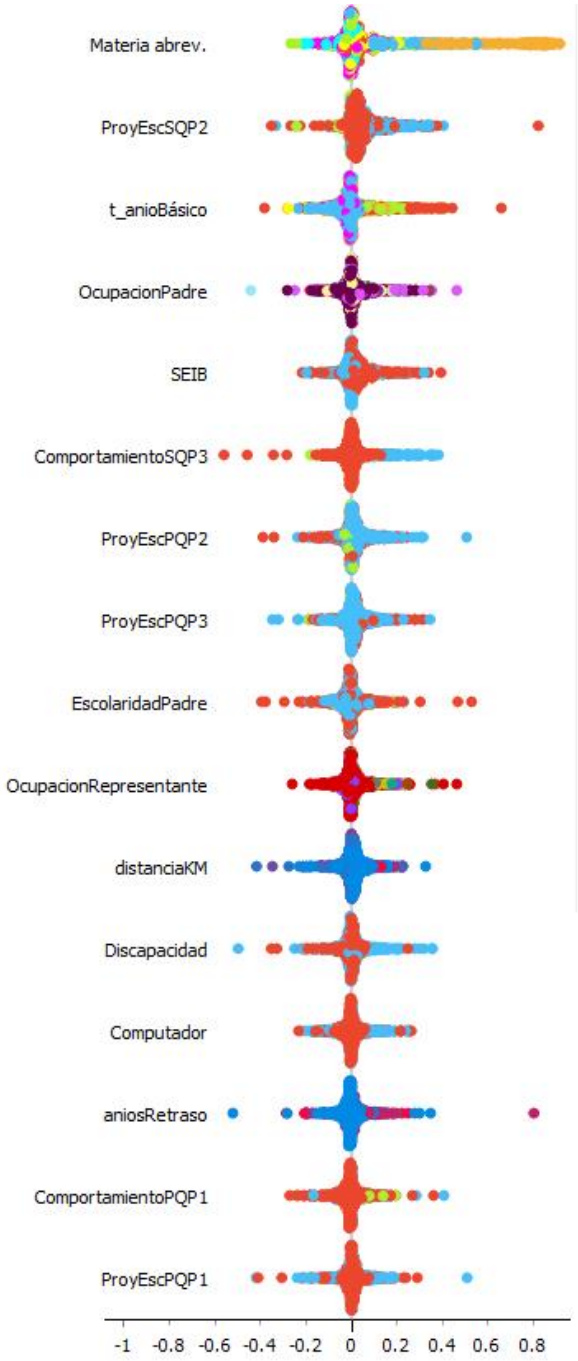

Figura 2. Importancia de las características en la clasificación de promedios DAR.

Los promedios de la clase AAR son los más suscitados en la muestra de datos, representan a los alumnos que aprueban con un poco más de lo justo. Para la clasificación de los promedios de esta clase impactan más la materia, proyectos escolares, ocupación del padre, año básico, tener discapacidad o no, género, dificultad auto reportada, número de hermanos, estructura familiar, distancia casa-escuela y ocupación de la madre[3].

Para la clasificación de los promedios de la clase PAAR impactan más la materia, proyectos escolares, ocupación del padre, tener discapacidad o no, año básico, dificultad auto reportada, género, distancia casa-escuela, escolaridad de la madre, ingresos familiares y escolaridad del padre. Esta clase representa a alumnos que por lo menos una vez alcanzan promedios parciales por debajo del mínimo necesario para aprobar el curso[4].

Para la clasificación de los promedios de la clase NAAR impactan más los proyectos escolares, la materia, distancia casa-escuela, año básico, escolaridad del padre, ocupación del padre, número de hermanos, tener discapacidad o no, ingresos familiares y género. Nótese que se puede entender que las habilidades sociales y la materia (usualmente matemática y lenguaje) son los factores que predominan en la obtención de este tipo de promedios. Esta clase representa a alumnos que por lo menos una vez alcanzan promedios parciales de los más bajos posibles[5].

Figuras como la **2** son de gran importancia porque clasifican a las características en orden decreciente. También muestran el grado de influencia de la característica mediante colores, con el rojo indicando un fuerte impacto (Eje Y máximo) y el azul indicando un bajo impacto (Eje Y Mínimo), así como la asociación positiva y negativa de la característica con el promedio que hace de variable objetivo. En el eje X, los gráficos de resumen se refieren a la interpretabilidad global. Entre otros, del gráfico se obtiene que la materia es la característica más relevante para clasificar a estudiantes con promedios DAR. Los alumnos con más años de retrasos tienen más posibilidades de obtener promedios DAR.

4. Discusión

El rendimiento académico de los estudiantes es crucial para su éxito futuro y un tema de gran preocupación para los académicos de todo el mundo [22]. Con base en los ítems precedentes, se buscó potenciar algunos aspectos, por ejemplo, el uso de métodos de aprendizaje en conjunto o ensamblados, que a su vez se apoyan en otros muy recurridos como lo son los árboles de decisión CART, es decir, se basan en la potenciación del gradiente y es de esperar que al tratarse de métodos relativamente nuevos e iterativos aporten ventajas respecto de sus antecesores. CatBoost logra mejores resultados cuando se emplea con características de entrada mayormente de tipo categórico porque justamente fue esa una de las razones de su surgimiento [8]. Los resultados sugieren que el modelo evaluado con un muestreo estratificado tiene una tasa de precisión en torno al 91%, para alcanzar dicho valor también se abordó el problema del desequilibrio de las clases subrepresentadas y se optimizó los hiperparámetros, especialmente la tasa de aprendizaje con un valor de 0.50 sobre 1.

Para ilustrar cómo CatBoost predice con el 91% de Exactitud se utilizó los valores SHAP. El valor de forma y las visualizaciones que lo acompañan ofrecen una visión del funcionamiento interno de CatBoost, aumentando la transparencia y utilidad para los administradores de escuelas e incluso otros niveles educativos. Se pretende que esta investigación ayude a comprender la problemática

multifactorial estudiada y proporcione información sobre modelos significativos para mejorar la educación en todo el mundo.

Se reconoce que los valores SHAP deben de pasar por todas las combinaciones concebibles de características para que el modelo sea interpretable y que cuando el número de características sea alto, el número de combinaciones posibles también lo será, lo que da lugar a una gran cantidad de cálculos con su consecuente complejidad temporal.

Conclusiones

En este artículo, se ha desarrollado y presentado un modelo de clasificación del rendimiento de los alumnos a partir de datos relacionados con sus calificaciones e información socioeconómica. Estos datos se obtuvieron de dos escuelas ecuatorianas como se ha descrito anteriormente.

Se enfatizó en la interpretabilidad, entendimiento y lógica subyacente del modelo para propiciar la generación de acciones de mejora del rendimiento académico. Los valores SHAP del modelo explican el 91% de Exactitud como una lista jerarquizada de características que expresan a los siguientes aspectos socioeconómicos como incidentes en el rendimiento escolar, figuran la asignatura, las habilidades sociales de los niños, la edad expresada por medio del año que cursa el alumno, ocupación del padre y en consecuencia el ingreso familiar, género, estructura familiar, ocupación de la madre y número de hermanos.

La métrica de la Exactitud que se ha obtenido es estable dado que se recurrió al uso de 100 estimadores CART con el método de aprendizaje en conjunto CatBoost, método que aprovecha las ventajas del Boosting o potenciación del gradiente.

CatBoost logra mejores resultados cuando se emplea con características de entrada mayormente de tipo categórico como es el caso de este estudio, en tanto que las visualizaciones de SHAP permiten transparentar el funcionamiento interno de los modelos de predicción y, por tanto, permite a los educadores identificar a los niños en situación de riesgo y realizar intervenciones eficaces desde los actores a los que le compete.

Limitaciones y futuro

Esta investigación presenta varios inconvenientes que conviene mencionar, como lo es la muestra de apenas dos escuelas. Un estudio con más datos podría reportar resultados más concluyentes

En la actualidad muchos de los modelos de aprendizaje automático son percibidos por los usuarios como cajas negras, sin embargo, en esta investigación se intentó presentar resultados interpretables, porque la interpretabilidad es importante porque no todos los usuarios de este tipo de sistemas son de formación en estadística o de ciencia de datos. Entonces resulta imperioso generar soluciones de aprendizaje automático interactivas e interpretativas que con base en normativas emergentes como es el caso del Reglamento General de Protección de Datos (GDPR) de la Unión Europea, transparenten a los usuarios la razón de sus decisiones.

Si bien los datos analizados, se han obtenido en formato tabular siguiendo un detallado proceso de extracción y preparación, se puede ampliar el estudio a otros datos como los de tipo psicológico, estilos de aprendizaje estudiantil, autoeficacia (de la que se ha tomado en consideración la dificultad auto reportada en las asignaturas), metas de logro, motivación, intereses u otros.

Agradecimientos

Se expresa nuestra gratitud con las dos escuelas ecuatorianas que nos dieron apertura para el uso responsable de sus datos, previo acuerdo de anonimización de información confidencial. Este estudio se financió parcialmente por medio del Departamento de Investigación de la Universidad Laica Eloy Alfaro de Manabí.

Notas

[1] Desde el siguiente enlace se proporciona una tabla que lista las columnas del conjunto de datos, su tipo y una breve descripción https://tinyurl.com/4w9wp6zr

[2] Desde el siguiente enlace se proporciona se proporciona una ficha de las nuevas características predictoras y de respuestas agregadas para el análisis https://tinyurl.com/29txpcf5

[3] Desde este enlace se muestra la figura de valores SHAP correspondientes con la clase AAR https://tinyurl.com/46rmy523

[4] Desde este enlace se muestra la figura de valores SHAP correspondientes con la clase PAAR https://tinyurl.com/ym8357d5

[5] Desde este enlace se muestra la figura de valores SHAP correspondientes con la clase NAAR https://tinyurl.com/54sfnxm8

Referencias

[1] M. Soncin y M. Cannistrà, "Data analytics in education: are schools on the long and winding road?", *Qualitative Research in Accounting & Management*, vol. 19, no. 3, 2022, doi: https://doi.org/10.1108/QRAM-04-2021-0058.

[2] K. Schildkamp, "Data-based decision-making for school improvement: Research insights and gaps", *Educational Research*, vol. 61, no. 3, Art. no. 3, jul. 2019, doi: https://doi.org/10.1080/00131881.2019.1625716 .

[3] J. I. Pincay-Ponce, J. S. Herrera-Tapia, J. Terranova-Ruiz, M. Cruz-Felipe, J. C. Sendón-Varela, y L. Fernández-

Capestany, "Minería de datos educativos: Incidencia de factores socioeconómicos en el aprovechamiento escolar", *Revista Ibérica de Sistemas e Tecnologias de Informação*, no. E49, 2022.

[4] J. I. Pincay-Ponce, "Reflexiones sobre la accesibilidad web para el contenido educativo en los sistemas de administración de aprendizaje", *REFCalE: Revista Electrónica Formación y Calidad Educativa. ISSN 1390-9010*, vol. 6, no. 1, pp. 193-206, 2018.

[5] J. I. Pincay-Ponce, J. S. Herrera-Tapia, J. Terranova-Ruiz, M. Cruz-Felipe, J. C. Sendón-Varela, y L. Fernández-Capestany, "Analítica de datos de factores socioeconómicos que inciden en el rendimiento escolar. Revisión sistemática", *Revista Ibérica de Sistemas e Tecnologias de Informação*, no. E52, Art. no. E52, 2023.

[6] F. Ofori, E. Maina, y R. Gitonga, "Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review", pp. 2616-3573, mar. 2020.

[7] J. Brownlee, *Ensemble Learning Algorithms With Python*, 1.11. Machine Learning Mastery, 2021.

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, y A. Gulin, "CatBoost: unbiased boosting with categorical features", no. arXiv:1706.09516. arXiv, 20 de enero de 2019. Accedido: 22 de noviembre de 2022. [En línea]. Disponible en: http://arxiv.org/abs/1706.09516

[9] X. Xiang, S. Duan, H. Pan, P. Han, J. Cao, y C. Liu, "From One-hot Encoding to Privacy-preserving Synthetic Electronic Health Records Embedding", en *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies*, Guangzhou China: ACM, dic. 2020, pp. 407-413. doi: https://doi.org/10.1145/3444370.3444605 .

[10] A. Joshi, P. Saggar, R. Jain, M. Sharma, D. Gupta, y A. Khanna, "CatBoost — An Ensemble Machine Learning Model for Prediction and Classification of Student Academic Performance", *Adv. Data Sci. Adapt. Data Anal.*, vol. 13, no. 03n04, p. 2141002, jul. 2021, doi: https://doi.org/10.1142/S2424922X21410023.

[11] Z. Mingyu, W. Sutong, W. Yanzhang, y W. Dujuan, "An interpretable prediction method for university student academic crisis warning", *Complex Intell. Syst.*, vol. 8, no. 1, pp. 323-336, feb. 2022, doi: https://doi.org/10.1007/s40747-021-00383-0.

[12] S. Lundberg y S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", 2017, doi: https://doi.org/10.48550/ARXIV.1705.07874.

[13] F. Grina, Z. Elouedi, y E. Lefevre, "Learning from Imbalanced Data Using an Evidential Undersampling-Based Ensemble", en *Scalable Uncertainty Management*, F. Dupin de Saint-Cyr, M. Öztürk-Escoffier, y N. Potyka, Eds., en Lecture Notes in Computer Science, vol. 13562. Cham: Springer International Publishing, 2022, pp. 235-248. doi: https://doi.org/10.1007/978-3-031-18843-5_16.

[14] G. Lemaître, F. Nogueira, y C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning", *Journal of Machine Learning Research*, vol. 18, no. 17, Art. no. 17, 2017.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.

[16] A. Fernández, S. Garcia, F. Herrera, y N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary", *Journal of artificial intelligence research*, vol. 61, pp. 863-905, 2018.

[17] Ministerio de Educación, "Instructivo para la aplicación de la evaluación estudiantil (actualizado a julio 2016)". Ministerio de Educación, 2016. [En línea]. Disponible en: https://tinyurl.com/ycc6tdvz

[18] Ministerio de Educación del Ecuador, "Instructivo para la aplicación de la evaluación estudiantil", Ministerio de Educación del Ecuador, Quito, Ecuador, 2013.

[19] J. Liu, G. Liang, K. D. Siegmund, y J. P. Lewinger, "Data integration by multi-tuning parameter elastic net regression", *BMC Bioinformatics*, vol. 19, no. 1, Art. no. 1, dic. 2018, doi: https://doi.org/10.1186/s12859-018-2401-1.

[20] S. Mukhopadhyay, *Advanced Data Analytics Using Python*. Berkeley, CA: Apress, 2018. doi: https://doi.org/10.1007/978-1-4842-3450-1.

[21] Ministerio de Educación del Ecuador, "Proyectos escolares. Instructivo", Ministerio de Educación del Ecuador, Quito, Ecuador, 2016.

[22] UNESCO, "Resultados de logros de aprendizaje y factores asociados del Estudio Regional Comparativo y Explicativo (ERCE 2019)", 2021. https://www.unesco.org/es/articles/resultados-de-logros-de-aprendizaje-y-factores-asociados-del-estudio-regional-comparativo-y## (accedido 19 de octubre de 2022).

Información de Contacto de los Autores:

Jorge Iván Pincay Ponce
Universidad Laica Eloy Alfaro de Manabí (ULEAM)
Vía a San Mateo Km 1.5 y Calle 12. Manta
Ecuador
jorge.pincay@uleam.edu.ec
https://orcid.org/0000-0003-4711-8850

Armando De Giusti
Universidad Nacional de La Plata (UNLP)
50 y 120 (1900). La Plata
Argentina
degiusti@lidi.info.unlp.edu.ar
https://orcid.org/0000-0002-6459-3592

Diana Alexandra Sánchez Andrade
Universidad de Guayaquil (UG)
Av. Delta S/N y Av. Kennedy. Guayaquil
Ecuador
diana.sancheza@ug.edu.ec
https://orcid.org/0000-0001-5154-6984

Juan Alberto Figueroa Suárez
Universidad Laica Eloy Alfaro de Manabí (ULEAM)
Vía a San Mateo Km 1.5 y Calle 12. Manta
Ecuador
juan.figueroa@uleam.edu.ec
https://orcid.org/0000-0002-5740-110X

Jorge Iván Pincay Ponce
Doctor (c) en Informática. Máster en Tecnologías de Información y de las Comunicaciones. Máster en Ingeniería de Software. Ingeniero en Sistemas. Docente en las carreras de Tecnologías de Información y de Software en la Universidad Laica Eloy Alfaro de Manabí.

Armando De Giusti
Investigador Principal del CONICET y Director del Instituto de Investigación en Informática LIDI de la Universidad Nacional de La Plata, Argentina. Especialista en Tecnología Informática Aplicada en Educación, Ingeniero en Telecomunicaciones y Calculista Científico.

Diana Alexandra Sánchez Andrade
Máster en Visualización de Datos. Ingeniera en Sistemas. Investigadora sobre ciencia de datos, tecnologías accesibles e ingeniería de software. Cumple funciones de soporte informático en la Universidad de Guayaquil.

Juan Alberto Figueroa Suárez
Máster en Gerencia Educativa. Máster en Ingeniería de Software. Especialista en Diseño Curricular. Analista de Sistemas. Docente de las asignaturas de Informática en la Universidad Laica Eloy Alfaro de Manabí.