Interpretación de los Resultados por Regiones de la Argentina del Operativo de Evaluación de la Calidad Educativa, Utilizando Modelos de Machine Learning

Andrés Francisco Farías[1], Germán Antonio Montejano[2],

Ana Gabriela Garis[3], Andrés Alejandro Farías[4]

Sebastián Javier Farías[5]

National University of La Rioja, La Rioja, Argentina[1, 4]
National University of San Luis, San Luis, Argentina[2, 3]
afarias665@yahoo.com.ar[1], gmonte@unsl.edu.ar[2],
agaris@gmail.com[3], andresaf86@hotmail.com[4]
sebajavfarias@gmail.com[5]

Abstract. El presente trabajo se basa en la utilización de el modelo de Machine Learning: Gradient Boosting Classifier de la Librería Sklearn , en las Pruebas de Evaluación Estandarizadas "Aprender" que se desarrolló en la Argentina, para medir el Desempeño en Lengua y el Desempeño en Matemáticas. Se propone realizar este abordaje con los datos de la Evaluación de Sexto Grado de la Escuela Primaria, de la Edición 2018 de esta Evaluación de Calidad Educativa. En la etapa la investigación se analizó solamente el Desempeño en Lengua y los resultados se exponen en este documento por regiones de la Argentina. Se realizó una preselección variables, usando la librería de Sklearn: SelectKBest y luego se adoptó un solo modelo de Machine Learning: Gradient mencionado. Posteriormente se ejecutaron los cálculos por país y sus regiones y se efectuaron comparaciones

Keywords: SelectKBest, Gradient Boosting Classifier, Dataframe, Desempeño Educativo

1 Introducción

El procedimiento consistió en analizar las Evaluaciones Estandarizadas de Calidad Educativa en Argentina - Operativo Nacional "Aprender" trabajando con algún modelo de Machine Learning, para determinar, las variables más significativas [1], [2].

Se trabajó con el dataframe del Operativo Nacional "Aprender" de 2018 realizado en Sexto Grado de la Escuela Primaria a nivel censal, que contenían muchas variables para medir el Desempeño en Lengua, que pasaron por un proceso de preselección por

medio de la Librería Sklearn el selector: SelectKBest, para determinar si era necesario descartar algunas de ellas.

Posteriormente se adoptó un modelo de Machine Learning: Gradient Boosting Classifier a las variables preseleccionadas [3]. Al dataframe se le aplico el modelo seleccionado y los resultados fueron tabulados por regiones de la Argentina y luego gaficados para su interpretación.

2 Evaluaciones Estandarizadas de Calidad Educativa en Argentina - Evaluación Nacional “Aprender"

El Operativo Nacional “Aprender” es un dispositivo de evaluación de aprendizajes que desde el año 2018, viene implementándose en la Argentina, tanto en el Nivel Primario como en el Nivel Secundario de la escolaridad obligatoria.

Uno de los antecedentes que tuvo esta Evaluación Aprender, son los Operativos Nacionales de Evaluación (ONE) que se pusieron en práctica entre 1993 y 2013.

Como se expresó anteriormente, desde el 2018 se viene implementando esta evaluación que aborda principalmente las áreas de Lengua y Matemática, aunque en algunas ediciones también se evaluaron áreas como Ciencias Sociales y Ciencias Naturales, Educación Ciudadana y Producción Escrita.

Además de las diferentes áreas y grados de escolaridad, la prueba presenta variaciones en cuanto a su cobertura, habiéndose realizado de manera censal en algunos casos y de manera muestral en otros.

De manera resumida se presenta una serie histórica de los operativos Aprender desde la edición del 2016 hasta la última aplicada en 2024.

Estas evaluaciones estandarizadas no solo permiten analizar los desempeños en Lengua y Matemática de los estudiantes, sino que, mediante la implementación de cuadernillos complementarios, recolectan información sobre los contextos donde estos aprendizajes se desarrollan. Entre estos factores que influyen en los aprendizajes se analizan las características de las familias, el contexto educativo y los atributos de cada estudiante, el nivel educativo de los padres, actividades que realizan los estudiantes fuera del horario escolar, bienes del hogar, trayectoria educativa, entre otras.

Estos factores condicionantes de los aprendizajes se analizan desde la primera edición del Aprender y según el propio Ministerio de Educación, el modelo más adecuado que debería usarse para identificar todos los factores condicionantes de los aprendizajes serían los siguientes: aprendizajes de los alumnos o del promedio del aula; factores familiares no observables de los alumnos; desarrollo infantil, variable que pretende captar, principalmente, la nutrición y la estimulación recibidas en la etapa preescolar; nivel económico-social del hogar; nivel educativo alcanzado por la madre y el padre; capital humano de los alumnos; capital de las escuelas, subdividido en físico, humano y social; efecto del aula, también llamado en este informe el “misterio del aula”, variable que intenta captar la eficacia de cada docente y de su relación con los alumnos; factores institucionales, que incluyen la organización escolar (por ejemplo, gestión estatal o privada [4], [5].

El procesamiento de esta evaluación se realiza de acuerdo a 4 niveles de desempeño y se analizan esos datos de Lengua y Matemática, de acuerdo a las variaciones que presentan los distintos factores asociados [6].

3 Procedimiento para la interpretación de los resultados por regiones

Para realizar la aplicación del modelo de Machine Learning sobre los datos del Operativo Nacional "Aprender", se siguió el siguiente procedimiento [7], [8]:

- Análisis de los datos
- Preselección de las variables del dataframe
- Aplicación de un Modelo de Machine Learning
- Tabulación de los resultados por regiones
- Gráficos comparativos
- Interpretación

3.1 Análisis de los datos

El dataframe correspondiente al Desempeño en Lengua contiene 562.214 filas, con muchos datos perdidos y valores negativos, los que fueron reemplazados por el método de rellenado con la mediana dentro de cada columna.

3.2 Preselección de variables del dataframe

Desempeño en Lengua. Para realizar la preselección se utilizó de la Librería Sklearn el selector: SelectKBest y se tomó como valor de comparación el Fisher_score, de las distintas variables [9], [10].

Table 1. Seleccción de variables, utilizando SelectKBest

Variables	Fisher-score
sector	13352
ap4	7331
ap17	4935
ap21a	4840
ap14	4634
ap15	4529
isocioa	4079
ap8	3634
ap9	3141
ap16	2346
ap20	1854
ap10	1646
ap1	1542

ap6	1528
ap12	1385
ambito	951
ap37	920
ap31	878
ap11	706
ap38	510

3.3 Aplicación del modelo de Machine Learning

Desempeño en Lengua. Se aplica el modelo de Machine Learning: Gradient Boosting Classifier [11], [12], sobre el dataframe correspondiente al Desempeño en Lengua, utilizando las variables preseleccionadas y se obtienen el listado e las variables según su importancia, en la Tabla 2.

Tabla 2. Desempeño en Lengua en todo el País, importancia de las variables

Variables	Descripción de variables	Importancia
sector	Sector de gestión	0,2870
ap20	¿Te va bien en tu clase de Lengua?	0,1049
ap4	¿Con cuántas personas vivís?	0,0978
isocioa	Indice socioeconómico del alumno	0,0956
ap8	Aproximadamente, ¿cuántos libros hay donde vivís?	0,0828
ap14	Además de asistir a la escuela, ¿ayudás a tus padres o familiares en su trabajo?	0,0733
ap21a	En tu opinión, ¿cómo leés?	0,0632
ap17	¿Repetiste de grado alguna vez?	0,0515
ap15	¿Trabajás fuera de tu casa para alguien que no sea parte de tu familia?	0,0371
ap9	¿Cuál es el máximo nivel educativo de tu mamá?	0,0230
ap31	¿Buscás información o conversás sobre estos temas en internet? (redes sociales, páginas web, foros, etc.)	0,0179
ap6	¿Cuántas habitaciones tiene el lugar donde vivís, sin contar la cocina y el baño?	0,0136
ap10	¿Cuál es el máximo nivel educativo de tu papá?	0,0125
ap16	¿Fuiste a jardín de infantes?	0,0120
ap11	¿Tu mamá o tu papá pertenecen a pueblos indígenas o son descendientes de pueblos indígenas?	0,0089
ambito	Ámbito	0,0071
ap12	En tu casa, ¿hablan alguna lengua indígena?	0,0052
ap1	¿Cuántos años tenés?	0,0047
ap37	¿Tuviste que faltar a la escuela para acompañar a tu familia por traslados por razones de trabajo?	0,0011
ap38	¿Cuántos días faltaste a clase por esta razón?	0,0009

3.4 Tabulación de los resultados por regiones

Regiones de la Argentina. Las provincias de la Argentina, se agrupan formando reiones, en Tabla 3.

Tabla 3. Regiones de la Argentina

Región	Provincias
Cuyo	Mendoza, San Juan, San Luis
NOA	La Rioja, Catamarca, Tucumán, Salta, Jujuy, Santiago del Estero
NEA	Formosa, Chaco, Misiones, Corrientes
PAMPEANA	La Pampa, Buenos Aires, Córdoba, Santa Fe, Entre Ríos
PATAGÓNICA	Neuquén, Río Negro, Chubut, Santa Cruz, Tierra del Fuego

Desempeño en Lengua. Se repite el procedimiento anterior, en todas las regiones del País, en Tabla 4.

Tabla 4. Desempeño en Lengua, importancia de las variables por regiones del País

Variables	PAÍS	NOA	NEA	CUYO	PAMPEANA	PATAGÓNICA
sector	0,2870	0,2558	0,1916	0,1953	0,2815	0,1448
ap20	0,1049	0,1098	0,0742	0,1016	0,1124	0,1318
ap4	0,0978	0,0777	0,0922	0,1246	0,0940	0,0962
isocioa	0,0956	0,0752	0,0985	0,0822	0,0883	0,0589
ap8	0,0828	0,0676	0,0577	0,0861	0,0943	0,1079
ap14	0,0733	0,0950	0,0956	0,0604	0,0698	0,0836
ap21a	0,0632	0,0533	0,0484	0,0757	0,0716	0,1181
ap17	0,0515	0,0418	0,0509	0,1044	0,0454	0,0327
ap15	0,0371	0,0650	0,0602	0,0258	0,0323	0,0295
ap9	0,0230	0,0379	0,0534	0,0255	0,0167	0,0500
ap31	0,0179	0,0189	0,0176	0,0194	0,0152	0,0157
ap6	0,0136	0,0143	0,0186	0,0165	0,0261	0,0300
ap10	0,0125	0,0200	0,0274	0,0138	0,0096	0,0300
ap16	0,0120	0,0115	0,0129	0,0178	0,0136	0,0176
ap11	0,0089	0,0164	0,0331	0,0093	0,0073	0,0134
ambito	0,0071	0,0153	0,0422	0,0044	0,0069	0,0039
ap12	0,0052	0,0069	0,0039	0,0131	0,0088	0,0153
ap1	0,0047	0,0101	0,0130	0,0120	0,0051	0,0128
ap37	0,0011	0,0051	0,0055	0,0067	0,0007	0,0025
ap38	0,0009	0,0022	0,0030	0,0054	0,0004	0,0051

3.5 Gráficos comparativos

Con los datos de la Tabla 4, se realizan gráficos comparativos de cada región con el País para apreciar mejor las posibles diferencias.

Región Cuyo: Se observan diferencias en las primeras variables con respecto al total de País.

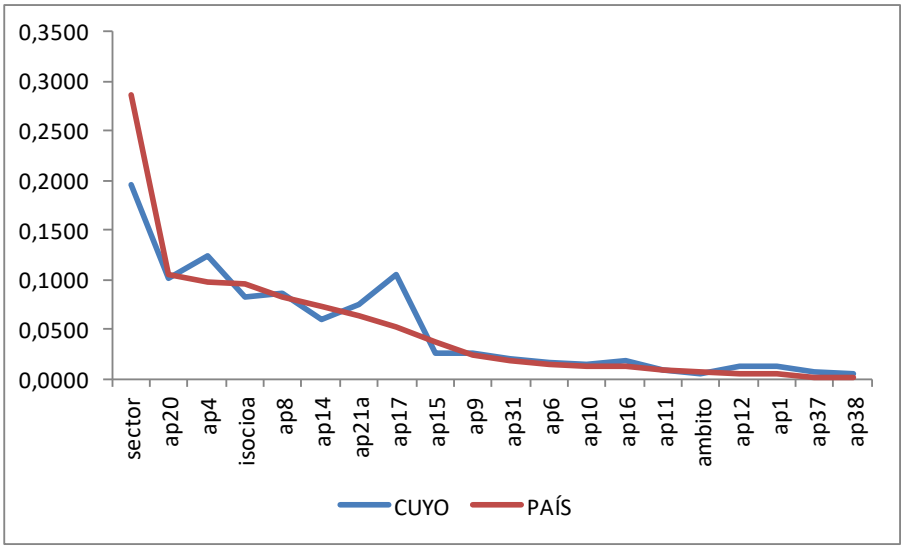

Fig. 1. Comparación de Cuyo con el País

Región NOA: Se observan diferencias leves con respecto al total de País.

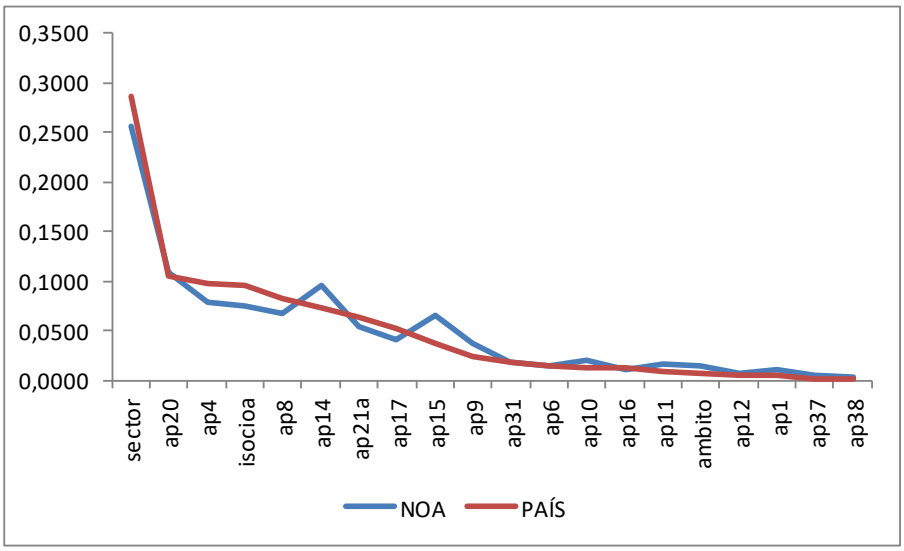

Fig. 2. Comparación de NOA con el País

Región NEA: Se observan diferencias con respecto al total de País.

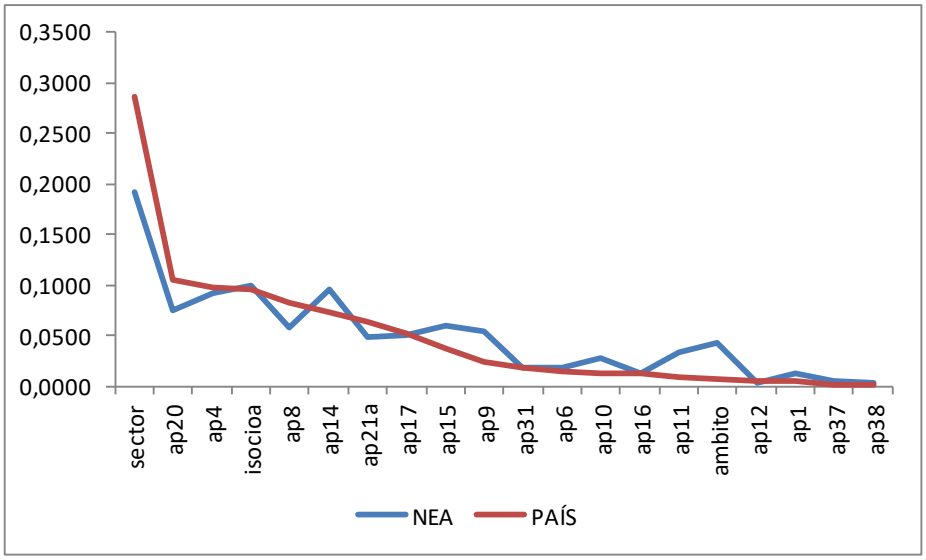

Fig. 3. Comparación de NEA con el País

Región PAMPEANA: Se observan coincidencias con respecto al total de País.

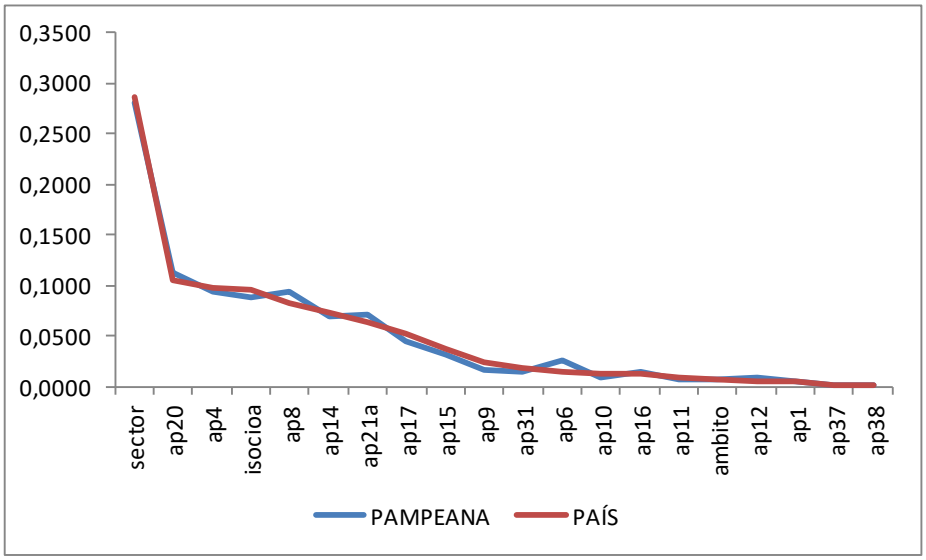

Fig. 4. Comparación de la Región Pampeana con el País

Región PATAGÓNICA: Se observan diferencias con respecto al total de País.

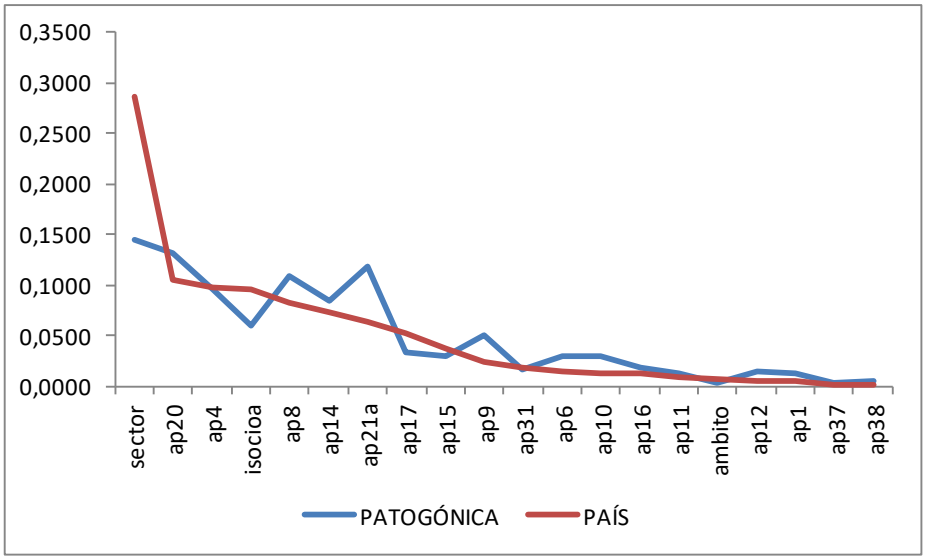

Fig. 5. Comparación de la Patagonia con el País

Comparación se todas las regiones con el País: Superposición de todos los gráficos

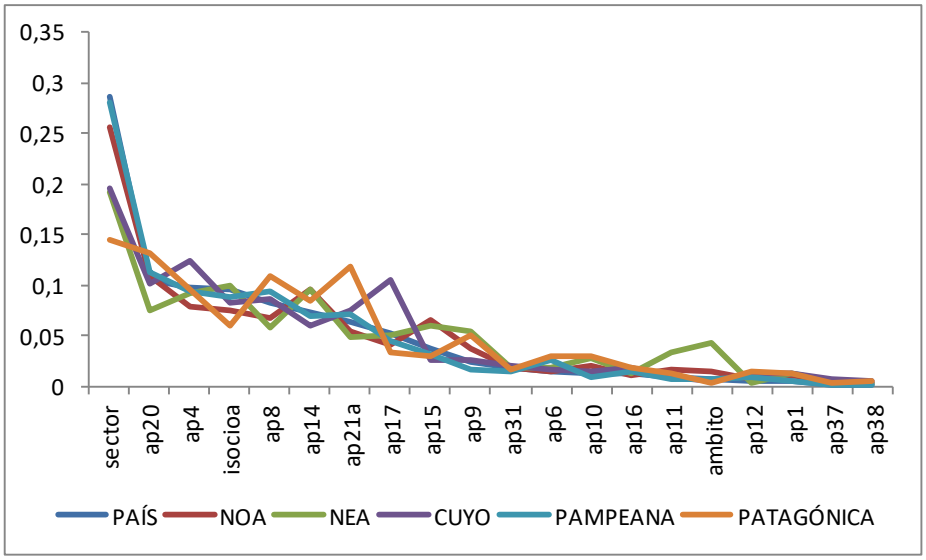

Fig. 7. Comparación de Regiones con el País

3.6 Interpretación

De la Tabla 2 y de los gráficos comparativos se aprecia que hay un grupo de variables significativas que ocupan los primeros lugares a nivel país y en las distintas regiones, ellas están indicadas en Tabla 5:

Tabla 5. Desempeño en Lengua, Primeras variables según su importancia

Variables	Descripción de variables
sector	Sector de gestión
ap20	¿Te va bien en tu clase de Lengua?
ap4	¿Con cuántas personas vivís?
isocioa	Indice socioeconómico del alumno
ap8	Aproximadamente, ¿cuántos libros hay donde vivís?
ap14	Además de asistir a la escuela, ¿ayudás a tus padres o familiares en su trabajo?
ap21a	En tu opinión, ¿cómo leés?
ap17	¿Repetiste de grado alguna vez?
ap15	¿Trabajás fuera de tu casa para alguien que no sea parte de tu familia?
ap9	¿Cuál es el máximo nivel educativo de tu mamá?

Hay diferencias entre las regiones y la regiones con respecto al País, pero se mantiene el predominio de las variables mencionadas.

Situación esperada y coincidente con los análisis estadísticos realizados por el Ministerio.

4 Conclusiones

Se ha cumplido con éxito la premisa del presente trabajo, demostrar la viabilidad del uso de las herramientas del Análisis de Datos y Machine Learning en las pruebas de Evaluaciones Nacionales que manejan enormes volúmenes de datos.

Es destacable observar que procesando la valiosa información que nos entregan los programas de evaluaciones, se llega por medio de los modelos de Machine Learning a resultados que arrojan las mismas tendencias que los obtenidos con los análisis estadísticos realizados sobre estas mismas pruebas.

En definitiva, procedimientos más expeditivos que brindan la información necesaria y oportuna para tomar las mejores decisiones.

Referencias

[1] Llach, J. J., & Cornejo, M. (2018). *Factores condicionantes de los aprendizajes en la escuela primaria y media.* In LIII Reunión Anual de la Asociación Argentina de Economía Política (La Plata, 14 al 16 de noviembre de 2018).

[2] Ministerio de Educación, Cultura, Ciencia y Tecnología. (2018) *Factores Condicionantes de los Aprendizajes. Primaria y Secundaria. Ciudad Autónoma de Buenos Aires*, Argentina. Recuperado de

https://www.argentina.gob.ar/sites/default/files/factores_condicionantes_de_los_aprendizajes.pdf

[3] Abro, A. A., Khan, A. A., Talpur, M. S. H., Kayijuka, I., & Yaşar, E. (2021). Machine learning classifiers: a brief primer. *University of Sindh Journal of Information and Communication Technology*, *5*(2), 63-68

[4] Llach, J. J., Schumacher, F., & Llach, J. (2006). *La segregación social en la educación primaria argentina*. Buenos Aires: Granica.

[5] Ministerio de Educación y Deportes. Secretaría de Evaluación Educativa (2016) *Aprender 20216. Informe de Resultados*. Recuperado de https://www.argentina.gob.ar/sites/default/files/reporte_nacional.pdf

[6] Adrogué, C. (2014). *Calidad y equidad de la educación primaria pública argentina*. Pilquen-Sección Psicopedagogía, 11(1), 12.

[7] Chakrapani, P., & Chitradevi, D. (2022, April). *Academic performance prediction using machine learning: A comprehensive & systematic review*. In 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC*)* (pp. 335-340). IEEE.

[8] Adedeji, O. B., Olayinka, O. O., Adebare, A., Idowu, P. A., Ayoade, A. O., & Ademola, A. O. (2025). *A Machine Learning-Based Predictive Model for the Classification of Academic Performance of Students*. University of Ibadan Journal of Science and Logics in ICT Research, 13(1), 69-79.

[9] Ververidis, D., & Kotropoulos, C. (2005, September). *Sequential forward feature selection with low computational* cost. In 2005 13th European Signal Processing Conference (pp. 1-4). IEEE.

[10] Kaur, A., Guleria, K., & Trivedi, N. K. (2021, March). Feature selection in machine learning: Methods and comparison. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 789-795). IEEE.

[11] Konstantinov, A. V., & Utkin, L. V. (2021). *Interpretable machine learning with an ensemble of gradient boosting machines.* Knowledge-Based Systems, 222, 106993.

[12] Kunapuli, G. (2023). *Ensemble methods for machine learning*. Simon and Schuster.