



# **COLEGIO DE POSTGRADUADOS**

---

**INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS**

**CAMPUS MONTECILLO**

**POSTGRADO EN SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA**

**CÓMPUTO APLICADO**

## **APLICACIÓN DE APRENDIZAJE AUTOMÁTICO PARA EVALUAR EL RENDIMIENTO ESCOLAR A PARTIR DE LA PRUEBA ENLACE**

**MIGUEL ÁNGEL MORALES HERNÁNDEZ**

**T E S I S**

**PRESENTADA COMO REQUISITO PARCIAL  
PARA OBTENER EL GRADO DE:**

**MAESTRO EN CIENCIAS**

**MONTECILLO, TEXCOCO, ESTADO DE MÉXICO**

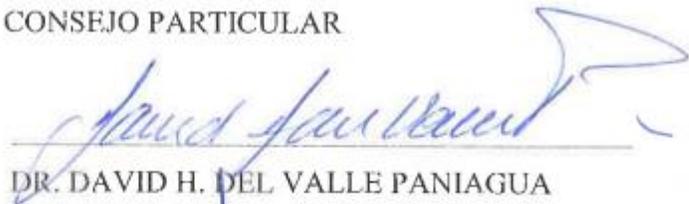
**2021**

La presente tesis titulada: **APLICACIÓN DE APRENDIZAJE AUTOMÁTICO PARA EVALUAR EL RENDIMIENTO ESCOLAR A PARTIR DE LA PRUEBA ENLACE**, realizada por el alumno: **MIGUEL ÁNGEL MORALES HERNÁNDEZ**, bajo la dirección del Consejo Particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIAS  
SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA  
CÓMPUTO APLICADO

CONSEJO PARTICULAR

CONSEJERO

  
DR. DAVID H. DEL VALLE PANIAGUA

ASESOR

  
DR. JUAN MANUEL GONZALEZ CAMACHO

ASESOR

  
M.C. JOSÉ RAFAEL DURÁN MORENO

Montecillo, Texcoco, México, diciembre de 2021.

# **APLICACIÓN DE APRENDIZAJE AUTOMÁTICO PARA EVALUAR EL RENDIMIENTO ESCOLAR A PARTIR DE LA PRUEBA ENLACE**

Miguel Ángel Morales Hernández, M.C.

Colegio de Postgraduados, 2021

## **RESUMEN**

El logro académico de los estudiantes está relacionado con diferentes variables que lo afectan, algunas son de características personales de los alumnos, estructurales o de ubicación geográfica de las escuelas, características del entorno escolar entre otras. La capacidad para predecir el rendimiento académico de los estudiantes ofrece beneficios a los profesores, alumnos y a las escuelas, para poder generar programas estratégicos de prevención o acompañamiento de estudiantes con bajo desempeño, detección de alumnos con alta probabilidad de deserción e identificación de características de los estudiantes que permita un buen aprovechamiento. A través de algoritmos de aprendizaje automático, es posible clasificar a estudiantes de acuerdo con características determinadas, por ejemplo, la calificación que obtuvieron en una materia o todo su historial académico. En esta investigación se implementan tres algoritmos de aprendizaje automático supervisado, bosque aleatorio, red neuronal multicapa y potenciación del gradiente para clasificar a estudiantes de la prueba ENLACE en tres clases o niveles de logro, 0: insuficiente, 1: elemental, 2: bueno o excelente en las asignaturas de español y matemáticas de 2008 y 2011. Se consideraron 13 variables predictoras que pueden influir en la clasificación. Bosque aleatorio se aplica para seleccionar las variables más importantes que se incluirán en los modelos; red neuronal multicapa y potenciación del gradiente para clasificar a los alumnos de tres entidades, Tlaxcala, Guerrero y Nuevo León. Los resultados muestran que es importante conocer previamente la calificación de los alumnos en español para clasificarlos en matemáticas y la calificación en matemáticas para predecir su clasificación en español, esto como una medida general de capacidad. Con la red neuronal multicapa se obtienen mejores resultados de clasificación en todas las entidades para español y con el modelo de potenciación del gradiente se obtienen mejores precisiones de clasificación para matemáticas. Al comparar los niveles de logro por sexo en las tres entidades, se observa que en español y matemáticas la proporción de mujeres en los niveles de logro académico 1 y 2 es mayor a la de los hombres.

**Palabras clave:** Alumno, características escolares, clasificación, variable predictora.

# **MACHINE LEARNING APP FOR TO ASSESS SCHOOL PERFORMANCE FROM THE ENLACE TEST**

Miguel Angel Morales Hernández, M.C.

Colegio de Postgraduados, 2021

## **ABSTRACT**

The academic achievement of students is related to different variables that affect it, some personal characteristics of the students, structural or geographic location of the schools, characteristics of the school environment, among others. The ability to predict the academic performance of students offers benefits to teachers, students, and schools, to be able to generate strategic programs for the prevention or monitoring of students with low performance, detection of students with a high probability of dropping out and identification of characteristics of the students that allow good achievement. Through machine learning algorithms, it is possible to classify students according to certain characteristics, for example, the grade they obtained in a subject or their entire academic history. In this research, three supervised machine learning algorithms, random forest, multilayer neural network and gradient boosting are implemented to classify ENLACE test students into three classes or levels of achievement, 0: insufficient, 1: elementary, 2: good or excellent in spanish and mathematics subjects 2008 and 2011. Thirteen predictor variables that may influence the classification were considered. Random forest is applied to select the most important variables to be included in the models; Multilayer neural network and gradient boosting to classify students from three entities, Tlaxcala, Guerrero, and Nuevo León. The results show that it is important to previously know the qualification of the students in spanish to classify them in mathematics and the qualification in mathematics to predict their classification in spanish, this as a general measure of ability. With the multilayer neural network, better classification results are obtained in all entities for spanish and with the gradient boosting model, better classification accuracies are obtained for mathematics. When comparing achievement levels by sex in the three entities, it is observed that in spanish and mathematics the proportion of women in academic achievement levels 1 and 2 is higher than that of men.

**Keywords:** Student, school characteristics, classification, predictor variable.

## **AGRADECIMIENTOS**

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado para la realización de mis estudios de posgrado.

Al Colegio de Postgraduados Campus Montecillo por darme la oportunidad de pertenecer a su sistema educativo para estudiar la maestría en Cómputo Aplicado y brindarme los conocimientos necesarios en mi formación, a través de sus cursos de excelencia.

Al Dr. David H. del Valle Paniagua quien fue mi consejero durante mi estancia en la maestría, por su apoyo para la realización de la tesis.

Al Dr. Juan Manuel González Camacho por su orientación y observaciones. Por brindarme la confianza para llevar a buen término este trabajo.

A mi asesor y a mi sinodal: M. C. José Rafael Durán Moreno y Dra. Yolanda Fernández Ordoñez por sus comentarios y el apoyo durante la realización de esta tesis.

Al Dr. Héctor Robles Vásquez por la lectura del trabajo y su valiosa asesoría. Persona importante en mi formación profesional y laboral.

Al Dr. Gerardo Terrazas González, excelente profesor y amigo, y de quien siempre recibí apoyo para continuar mis estudios.

A mis padres, Carolina y Gabino, a ellos con mucho cariño y respeto por ayudarme a llegar hasta este punto de mi vida. Porque de ellos he aprendido que con humildad, trabajo, esfuerzo y dedicación se puede construir un mejor futuro.

A todos mis hermanos, especialmente a Mariana y Concha, con quienes conviví más y me acompañaron durante estos dos años.

A Domingo Martínez, gran amigo, quien fue un apoyo importante durante mi estancia en la maestría, por sus consejos y motivación. A German Sánchez, por acercarme nuevamente al mundo de la lectura y por su valiosa amistad.

## **DEDICATORIA**

Este grado se lo dedico con mucho cariño a mis padres y a toda mi gran familia.

A mis hermanos, amigos y todos aquellos que siempre me apoyaron durante mi estancia en la maestría.

## CONTENIDO

RESUMEN .....	iii
ABSTRACT .....	iv
LISTA DE FIGURAS .....	x
LISTA DE CUADROS .....	xii
INTRODUCCIÓN GENERAL .....	1
OBJETIVOS E HIPÓTESIS .....	3
Objetivo General.....	3
Objetivos específicos .....	3
Hipótesis .....	3
CAPÍTULO I. ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DEL LOGRO ACADÉMICO A PARTIR DE LA PRUEBA ENLACE EN TLAXCALA, MÉXICO .....	4
1.1 RESUMEN .....	4
1.2 ABSTRACT.....	6
1.3 INTRODUCCIÓN .....	8
1.4 MATERIALES Y MÉTODOS .....	10
1.4.1 Colecta y preparación de datos .....	10
1.4.2 Clasificadores de aprendizaje automático.....	14
1.4.2.1 Red neuronal multicapa.....	15
1.4.2.2 Clasificador potenciación del gradiente .....	17
1.4.2.3 Bosque aleatorio .....	17
1.4.3 Criterios de desempeño de los modelos de predicción .....	18
1.4.4 Entrenamiento y validación de los modelos.....	19
1.4.4.1 Selección de hiperparámetros óptimos.....	19
1.4.4.2 Evaluación del desempeño de los modelos de clasificación .....	20
1.4.5 Importancia relativa de las características de entrada .....	21
1.5 RESULTADOS Y DISCUSIÓN .....	22
1.5.1 Hiperparámetros óptimos .....	22
1.5.2 Desempeño de los clasificadores .....	22
1.5.3 Importancia relativa de características .....	26

1.5.4 Discusión.....	26
1.6 CONCLUSIONES.....	28
CAPÍTULO II. COMPARAR LA PREDICCIÓN DE LOS NIVELES DE LOGRO ACADÉMICO EN EL NIVEL BÁSICO EN LOS ESTADOS DE TLAXCALA, GUERRERO Y NUEVO LEÓN CON BASE EN ALGORITMOS DE APRENDIZAJE AUTOMÁTICO.....	29
2.1 RESUMEN .....	29
2.2 ABSTRACT.....	31
2.3 INTRODUCCIÓN .....	33
2.4 MATERIALES Y MÉTODOS .....	35
2.4.1 Colecta y preparación de datos .....	35
2.4.2 Análisis exploratorio de los datos .....	39
2.4.3 Clasificadores de aprendizaje automático.....	39
2.4.3.1 Red neuronal multicapa.....	40
2.4.3.2 Clasificador potenciación del gradiente .....	40
2.4.3.3 Bosque aleatorio .....	41
2.4.4 Criterios de desempeño de los modelos de predicción .....	42
2.4.5 Entrenamiento y validación de los modelos.....	43
2.4.5.1 Selección de hiperparámetros óptimos.....	44
2.4.5.2 Evaluación del desempeño de los modelos de clasificación .....	45
2.4.6 Importancia relativa de las características de entrada .....	45
2.5 RESULTADOS Y DISCUSIÓN .....	46
2.5.1 Análisis exploratorio .....	46
2.5.2 Hiperparámetros óptimos.....	48
2.5.3 Desempeño de los clasificadores .....	50
2.5.4 Comparación de los modelos de clasificación .....	54
2.5.5 Importancia relativa de características .....	55
2.5.6 Discusión.....	56
2.6 CONCLUSIONES .....	57
CONCLUSIONES GENERALES.....	58
LITERATURA CITADA .....	59

ANEXO .....	62
DESCRIPCIÓN DE ESTRUCTURA DEL PROGRAMA EN PYTHON.....	62
RECOLECCIÓN DE DATOS Y PROCESAMIENTO.....	63
ENTRENAMIENTO Y SELECCIÓN DE MODELOS.....	64

## LISTA DE FIGURAS

### CAPÍTULO I

- Figura 1.1. Niveles de logro académico por sexo para los cuatro conjuntos de datos, ESP2008, ESP2011, MAT2008 y MAT2011. Fuente: elaboración propia. .... 14
- Figura 1.2. Proceso general del aprendizaje automático supervisado. Fuente: elaboración propia con base en Raschka y Mirjalili 2017. .... 15
- Figura 1.3. Arquitectura de un clasificador de red neuronal multicapa (MLP). Fuente: Raschka y Mirjalili, 2017. .... 16
- Figura 1.4. Curvas para el modelo MLP-español 2011. A) B) curva ROC, B) curva de P-S. Fuente: Elaboración propia. .... 23
- Figura 1.5. Ejemplo de curvas para el modelo GB-matemáticas 2011. A) curva ROC B) curva P-S. Fuente: Elaboración propia. .... 25
- Figura 1.6. Importancia de seis variables con RF. A) Español 2008, B) Matemáticas 2008, fuente: Elaboración propia. .... 26
- Figura 1.7. Importancia de seis variables con RF. A) Español 2011, B) Matemáticas 2011, fuente: Elaboración propia. .... 26

### CAPÍTULO II

- Figura 2.1. Arquitectura de un clasificador de red neuronal multicapa (MLP). Fuente: Elaboración propia con base en González-Camacho et al., 2012. .... 40
- Figura 2.2. Construcción de un bosque aleatorio. Fuente: Elaboración propia. .... 41
- Figura 2.3. Matriz de confusión y métricas de desempeño. Fuente: elaboración propia. .... 42
- Figura 2.4. Nivel de logro de los estudiantes de Tlaxcala, Guerrero y Nuevo León por sexo. A) Español 2008, B) Español 2011. Fuente: Elaboración propia. .... 46
- Figura 2.5. Nivel de logro de los estudiantes de Tlaxcala, Guerrero y Nuevo León por sexo. A) Matemáticas 2008, B) Matemáticas 2011. Fuente: Elaboración propia. .... 46

Figura 2.6. Nivel de logro de los estudiantes de Tlaxcala, Guerrero y Nuevo León por condición de becario A) Español 2008, B) Español 2011. Fuente: Elaboración propia.....	47
Figura 2.7. Nivel de logro de los estudiantes de Tlaxcala, Guerrero y Nuevo León por condición de becario. A) Matemáticas 2008, B) Matemáticas 2011. Fuente: Elaboración propia.....	47
Figura 2.8. Curva ROC para el modelo GB en español 2011, Guerrero. Fuente: elaboración propia. ....	51
Figura 2.9. Curva ROC para el modelo GB en español 2011, Nuevo León. Fuente: elaboración propia. ....	51
Figura 2.10. Curva ROC para el modelo GB en matemáticas 2011, Guerrero. Fuente: elaboración propia. ....	53
Figura 2.11. Curva ROC para el modelo GB en matemáticas 2011, Nuevo León. Fuente: elaboración propia. ....	53
Figura 2.12. Comparación de precisión global para los modelos por entidad federativa. A) Tlaxcala, B) Guerrero, C) Nuevo León. Fuente: Elaboración propia.....	55
Figura 2.13. Importancia de seis variables con RF. Español 2011 A) Tlaxcala, B) Guerrero C) Nuevo León. Matemáticas 2011 D) Tlaxcala E), Guerrero F) Nuevo León. Fuente Elaboración propia.....	56

## LISTA DE CUADROS

### CAPÍTULO I

Cuadro 1.1. Intervalos de puntaje para determinar el nivel de logro académico en español y matemáticas. ....	10
Cuadro 1.2. Variables contextuales y localización geográfica seleccionadas del Estado de Tlaxcala, México.....	12
Cuadro 1.3. Descripción de los conjuntos datos analizados por asignatura, año y niveles de logro académico en el estado de Tlaxcala para un total de 11 036 registros de estudiantes.....	13
Cuadro 1.4. Número de muestras del conjunto de prueba por clase objetivo. ....	19
Cuadro 1.5. Intervalo de valores para la prueba de hiperparámetros. ....	20
Cuadro 1.6. Hiperparámetros óptimos para el clasificador red neuronal multicapa y potenciación del gradiente y cada conjunto de datos (asignatura y año). ....	22
Cuadro 1.7. Precisión global de clasificación de los modelos por año y materia evaluada. ....	22
Cuadro 1.8. Valores de P y F1, AUCROC y AUCP-S para cada modelo por año y clase en español. ....	24
Cuadro 1.9. Valores de P y F1, AUCROC y AUCP-S para cada modelo por año y clase en matemáticas.....	25

### CAPÍTULO II

Cuadro 2.1. Variables contextuales y localización geográfica seleccionadas del Estado de Tlaxcala, Guerrero y Nuevo León, México. ....	36
Cuadro 2.2. Frecuencia de los niveles de logro de desempeño en español y matemáticas en Tlaxcala para una población de 11 036 estudiantes vigentes en 2008-2011.....	37
Cuadro 2.3. Frecuencia de los niveles de logro de desempeño en español y matemáticas en Guerrero para una población de 32 382 estudiantes vigentes en 2008-2011.....	38

Cuadro 2.4. Frecuencia de los niveles de logro de desempeño en español y matemáticas en Nuevo León para una población de 40 935 estudiantes vigentes en 2008-2011. ....	38
Cuadro 2.5. Muestras de prueba por clase, Tlaxcala. ....	43
Cuadro 2.6. Muestras de prueba por clase, Guerrero. ....	43
Cuadro 2.7. Muestras de prueba por clase, Nuevo León. ....	44
Cuadro 2.8. Intervalo de valores para la prueba de hiperparámetros para Tlaxcala, Guerrero y Nuevo León. ....	44
Cuadro 2.9. Hiperparámetros óptimos para el clasificador red neuronal multicapa (MLP) y potenciación del gradiente (GB) y cada conjunto de datos (asignatura y año), Tlaxcala. ....	48
Cuadro 2.10. Hiperparámetros óptimos para el clasificador red neuronal multicapa (MLP) y potenciación del gradiente (GB) y cada conjunto de datos (asignatura y año), Guerrero. ....	49
Cuadro 2.11. Hiperparámetros óptimos para el clasificador red neuronal multicapa (MLP) y potenciación del gradiente (GB) y cada conjunto de datos (asignatura y año), Nuevo León. ....	49
Cuadro 2.12. Precisión global-PG de clasificación de los modelos por año y materia evaluada. ....	50
Cuadro 2.13. Valores de P y F1, AUCROC y AUCP-S para cada modelo por año y clase en español, Guerrero. ....	52
Cuadro 2.14. Valores de P y F1, AUCROC y AUCP-S para cada modelo por año y clase en español, Nuevo León. ....	52
Cuadro 2.15. Valores de P y F1, AUCROC y AUCP-S para cada modelo por año y clase en matemáticas, Guerrero. ....	54
Cuadro 2.16. Valores de P y F1, AUCROC y AUCP-S para cada modelo por año y clase en matemáticas, Nuevo León. ....	54

## INTRODUCCIÓN GENERAL

México empezó a usar las pruebas estandarizadas para medir el logro académico de los estudiantes en las dos últimas décadas. La prueba de Evaluación Nacional de Logro Académico en Centros Escolares (ENLACE) se aplicó de manera censal a partir de 2006 a todos los alumnos de tercero a sexto de primaria y de los tres años de secundaria. En 2008 se empezó a aplicar a los tres años de bachillerato. El referente de esta prueba es el currículo nacional (Martínez, 2015).

Los resultados de la prueba se midieron en una escala estandarizada que va de 200 a 800 puntos y de acuerdo con la dificultad de los reactivos se establecen cuatro niveles de logro 0: insuficiente, 1: elemental, 2: bueno, 3: excelente. La información que se generó de ENLACE fue de utilidad para los docentes y padres de familia. El docente pudo comparar los resultados de su escuela con la de otras escuelas de características similares (socioeconómicas o estructurales). Además de identificar contenidos curriculares que los estudiantes no dominan para tomar las acciones pertinentes (SEP, 2008). A la par de ENLACE se aplicaron cuestionarios a una muestra de alumnos, padres de familia, docentes y directores de las escuelas para conocer las características personales, entorno familiar, hábitos de lectura, características de la vivienda, infraestructura escolar, métodos de enseñanza, etcétera., con la finalidad de contar con información de factores asociados a los aprendizajes. Un estudio precursor realizado en Estados Unidos (Informe Coleman), con efectos importantes en la política educativa, concluyó que la familia y la influencia de los compañeros de la escuela son factores más importantes en la determinación del logro académico de los alumnos de este país. Los recursos escolares tienen un efecto moderado en los resultados de logro (De Hoyos *et al.*, 2012).

En México existen pocos trabajos sobre factores que influyen en el desempeño escolar. Álvarez *et al.*, (2007) usaron los resultados del Programa para la Evaluación Internacional de Alumnos (PISA) 2003 en matemáticas, ciencias y lectura considerando variables asociadas al estudiante referentes a indicadores socioeconómicos, características de la escuela y cuestiones institucionales (pedagogía estatal, influencia sindical, etcétera.). Concluyeron que la rendición de cuentas y la evaluación son necesarios para mejorar los resultados en PISA. Mayer y Serván (2009) usaron información de la Encuesta Nacional sobre Niveles de Vida

de los Hogares (ENNViH) para estimar una función de producción estocástica. Concluyen que variables relacionadas con la riqueza familiar y características de la localidad (rural, urbana, marginación, etcétera.) son importantes en el desempeño escolar.

La educación de calidad tiene un papel relevante dentro de las metas propuestas en los Objetivos de Desarrollo Sostenible (ODS), que son respaldadas por las Naciones Unidas y que cada país miembro debe garantizar a su población. Un factor importante es pronosticar el desempeño académico de los estudiantes, que se puede lograr utilizando las herramientas minería de datos para información educativa y a través de procedimientos de aprendizaje automático. Hussain y Qasim (2021) utilizan una base de datos histórica (características personales) de 90 000 estudiantes de la Board of Intermediate & Secondary Education Peshawar (organismo gubernamental de evaluación y realización de exámenes en los niveles intermedios y secundaria) para pronosticar sus calificaciones usando dos modelos de aprendizaje supervisado: árboles de decisión (DT) y vecino más cercano (KNN) y una combinación de cada modelo con algoritmos genéticos (DT+GA, KNN+GA). El mejor modelo fue DT+GA con 96.6% de precisión global de clasificación y un error cuadrado de 5.34 para la predicción.

En este trabajo se implementan en el lenguaje de programación Python dos algoritmos de aprendizaje automático supervisado para clasificación, una red neuronal multicapa (MLP) y un modelo de potenciación del gradiente (GB) para clasificar a los alumnos en tres niveles de logro o clases, 0: insuficiente, 1: elemental, 2: bueno o excelente en español y matemáticas de 2008 y 2011, considerando como variables predictoras: puntaje en español, puntaje en matemáticas, becario, turno, sexo, tipo de escuela, etcétera. La tesis se divide en dos capítulos: en el primero se relacionan variables contextuales a la clasificación de los alumnos en tres clases o niveles de logro para el estado de Tlaxcala, se ajustan dos algoritmos MLP y GB; y se presentan la importancia de las variables en los modelos. En el segundo capítulo se hace una comparación de los niveles de logro para la variable *sexo* y *becario* para Tlaxcala, Guerrero, Nuevo León. Se implementan los dos algoritmos de aprendizaje supervisado, se comparan la precisión global obtenida de los mejores modelos en cada entidad, las precisiones obtenidas en las tres clases, la importancia de variables y las métricas de desempeño.

## **OBJETIVOS E HIPÓTESIS**

### **Objetivo General**

- Implementar dos modelos de aprendizaje automático supervisado para el análisis de variables contextuales y su influencia en el rendimiento escolar de los alumnos de la prueba ENLACE en 2008 y 2011, como determinantes en la finalización de la educación.

### **Objetivos específicos**

- Implementar en lenguaje de programación Python los algoritmos de aprendizaje automático supervisado, red neuronal multicapa y potenciación del gradiente.
- Considerar variables contextuales, determinar su importancia en los algoritmos.
- Aplicar los algoritmos de aprendizaje automático a los datos de Tlaxcala, Guerrero y Nuevo León, entidades con características sociodemográficas y económicas diferenciadas.
- Contrastar el nivel de logro académico en los estados de Tlaxcala, Guerrero y Nuevo León con base en variables contextuales obtenidas de la prueba ENLACE.

### **Hipótesis**

- Las condiciones socioeconómicas de los alumnos, sus pares y de la escuela influyen en el rendimiento académico de los alumnos.

# **CAPÍTULO I. ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DEL LOGRO ACADÉMICO A PARTIR DE LA PRUEBA ENLACE EN TLAXCALA, MÉXICO**

## **1.1 RESUMEN**

Las condiciones socioeconómicas de los alumnos, de sus pares y de la escuela condicionan el grado de logro académico de los alumnos en los diferentes niveles educativos. Por ello, resulta de interés analizar el efecto de estos factores en los procesos de aprendizaje como medida para implementar acciones diferenciadas de atención a los alumnos o como ayuda a la mejora de políticas públicas educativas. En esta investigación se implementaron dos clasificadores de aprendizaje automático, una red neuronal multicapa (MLP, multilayer perceptron) y un modelo potenciación del gradiente (GB, gradient boosting) para predecir el grado de logro académico (0: insuficiente; 1: elemental; 2: bueno o excelente) en español y matemáticas de alumnos de educación básica, en dos etapas sexto de primaria (2008) y tercero secundaria (2011) con base en variables contextuales obtenidas de ENLACE del Estado de Tlaxcala, México. Se consideraron 13 variables de entrada. La importancia relativa de éstas se determinó con el clasificador bosque aleatorio (RF, random forest). Los clasificadores MLP y GB se entrenaron y probaron con un conjunto de datos de 11 036 registros de estudiantes que permanecieron en el sistema escolar de 2008 a 2011. Los modelos se entrenaron y probaron en predicción para 2008 y 2011. En español el clasificador MLP fue superior a GB con una precisión global de clasificación ( $PG$ ) de 70.1% en 2008 y 61.1% en 2011. GB obtuvo mejores resultados en matemáticas con  $PG = 68.8\%$  en 2008 y 63.5% en 2011. Se observó que el puntaje en español tiene una fuerte asociación con el grado de logro académico en matemáticas. MLP obtuvo mejor desempeño para clasificar los grados de logro académico 1 y 2 de español con la métrica área bajo la curva precisión-sensibilidad ( $AUC_{P-S}$ ) con valores de 0.72 y 0.75. GB obtuvo mejor desempeño en matemáticas (sexto de primaria) para clasificar los grados de logro 1 y 2, con  $AUC_{P-S} = 0.72$  y 0.70 respectivamente. GB obtuvo buen desempeño para clasificar el grado de logro académico 0 en 2011, con  $AUC_{P-S}$  de 0.70 en español y de 0.82 en matemáticas. Los puntajes en español y matemáticas tuvieron mayor importancia relativa con respecto a los factores contextuales considerados como sexo, beca, turno de la escuela, etcétera. Esta importancia radica en el hecho de poder asociar variables que pueden influir en la clasificación del nivel de logro de los estudiantes y

que permitan señalar a los encargados de la política educativa enfocar estrategias para su atención oportuna.

De los datos originales de la prueba ENLACE para Tlaxcala, al analizar la distribución de hombres y mujeres por clases, hay más mujeres en las clases 1 y 2 en las dos materias evaluadas para ambos años. Y los hombres tienden a ser mayores en la clase 0 que representa en nivel de logro insuficiente. Así mismo, de 2008 a 2011 hay un aumento en el número de estudiantes en las clases 0 y 1 y una disminución en la clase 2. Estos movimientos se dan principalmente de alumnos que estaban en la clase 1 y pasaron a la clase 0, es decir, alumnos que son malos, en general siguen siendo malos.

**Palabras clave:** Aprendizaje supervisado, árboles de decisión, contexto escolar, redes neuronales artificiales, validación cruzada.

## 1.2 ABSTRACT

The socioeconomic conditions of the students, their peers and the school determine the degree of academic achievement of the students at the different educational levels. For this reason, it is of interest to analyze the effect of these factors on learning processes as a measure to implement differentiated actions of attention to students or as an aid to the improvement of educational public policies. In this research, two machine learning classifiers were implemented, a multilayer neural network (MLP, multilayer neural network) and a gradient boosting model (GB, gradient boosting) to predict the degree of academic achievement (0: insufficient; 1: elementary; 2 : good or excellent) in spanish and mathematics of basic education students, in two stages, sixth primary (2008) and third secondary (2011) based on contextual variables obtained from ENLACE test of the State of Tlaxcala, México. Thirteen input variables were considered. The relative importance of these was determined with the random forest (RF) classifier. The MLP and GB classifiers were trained and tested with a data set of 11 036 records of students who remained in the school system from 2008 to 2011. The models were trained and tested in prediction for 2008 and 2011. In spanish the MLP classifier was superior a GB with a global classification accuracy ( $PG$ ) of 70.1% in 2008 and 61.1% in 2011. GB obtained better results in mathematics with  $PG = 68.8\%$  in 2008 and 63.5% in 2011. It was observed that the score in spanish has a strong association with the degree of academic achievement in mathematics. MLP obtained better performance to classify achievement levels 1 and 2 of spanish with the metric area under the Precision-Sensitivity curve ( $AUC_{P-S}$ ) with values of 0.72 and 0.75. GB obtained better performance in mathematics (sixth grade) to classify achievement levels 1 and 2, with  $AUC_{P-S} = 0.72$  and 0.70 respectively. GB performed well for classify level 0 in 2011, with  $AUC_{P-S}$  of 0.70 in spanish and 0.82 in mathematics. Scores in spanish and mathematics have greater relative importance with respect to contextual factors considered as sex, scholarship, school shift, and so on. This importance lies in the fact of being able to associate variables that can influence the classification of the level of achievement of students and that allow those in charge of educational policy to focus strategies for their timely attention.

From the original data of the ENLACE test for Tlaxcala, when analyzing the distribution of men and women by classes, there are more women in classes 1 and 2 in the two subjects evaluated for both years. And men tend to be older in class 0, which represents insufficient

achievement. Likewise, from 2008 to 2011 there is an increase in the number of students in classes 0 and 1 and a decrease in class 2. These movements mainly occur in students who were in class 1 and went to class 0, that is in other words, students who are bad will generally remain bad.

**Key words:** Supervised learning, decision trees, school context, artificial neural networks, cross validation.

### 1.3 INTRODUCCIÓN

La evaluación del aprendizaje de los alumnos mediante pruebas a gran escala (estatal o nacional) permite obtener información sobre los grados de logro académico y variables contextuales asociadas. La Organización para la Cooperación y el Desarrollo Económicos (OCDE, 2005) encontró evidencias de cómo se relacionan factores como el contexto escolar, insumos y procesos escolares con el proceso de aprendizaje de los alumnos. En México, la Secretaría de Educación Pública (SEP) cuenta con bases de datos de alumnos que se matriculan anualmente en cada nivel educativo y resultados de las pruebas que se aplican a nivel nacional como los Exámenes de Calidad y Logro Educativo (EXCALE), ENLACE o internacional como el Programa Internacional de Evaluación de los Alumnos (PISA) (INEE, 2019).

La SEP implementó desde 2006 la prueba ENLACE de manera censal a alumnos de educación básica y bachillerato para evaluar el rendimiento escolar en español y matemáticas (SEP, 2008). Sin embargo, se cuenta con pocos estudios sobre la importancia de variables de tipo contextual en el logro académico de los estudiantes.

En la literatura se documentan aplicaciones de los modelos de aprendizaje automático para analizar el rendimiento escolar. Đambić, Krajcar y Bele (2016) utilizaron un modelo de regresión logística para la detección temprana de estudiantes con problemas en su rendimiento, en un curso de computación. El modelo obtuvo un error de clasificación de 19.0%. Rai *et al.*, (2020) utilizaron dos modelos de clasificación, bosques aleatorios y máquinas de soporte vectorial para predecir el rendimiento escolar de un grupo de estudiantes universitarios, con base en variables de entrada como: sexo, horas de estudio, porcentaje de asistencia a clases, ingreso mensual de la familia. El modelo bosque aleatorio obtuvo una precisión global de clasificación de 94.0% y la máquina de soporte vectorial 79.0%. Altabrawee, Osama y Qaisir (2019), aplicaron cuatro técnicas de aprendizaje automático red neuronal, decisión Bayesiana, árboles de decisión y regresión logística para predecir el rendimiento de los estudiantes en un curso de computación, con base en el efecto del uso del internet como medio de aprendizaje; algunas variables consideradas fueron el tiempo empleado en redes sociales, horas de estudio, sexo, educación de los miembros de la familia.

La red neuronal alcanzó el mejor desempeño con una precisión de 77.0% y área bajo la curva  $AUC_{ROC}$  de 0.807.

En México se han realizado trabajos relacionados con el efecto de factores externos e internos a la escuela en el logro académico de los estudiantes. Fernández (2003) utilizó índices de capital familiar global (nivel educativo de la madre, equipamiento de confort en la vivienda y disposición de libros y computadora), de contexto sociocultural de la escuela y de clima organizacional; se aplicó análisis jerárquico para explicar el aprendizaje en español y en matemáticas. Un aumento en el índice de capital familiar global incide en el incremento del resultado de español y matemáticas. Sin embargo, cuando se generalizan las carencias de la vivienda (analfabetismo, bajos ingresos), los resultados de los alumnos son bajos.

En esta investigación se implementaron dos clasificadores de aprendizaje automático una red neuronal multicapa y un algoritmo de ensamble potenciación del gradiente para predecir el grado de logro académico (0: insuficiente, 1: elemental, 2 : bueno o excelente) en español y matemáticas a alumnos de educación básica en dos etapas de su formación, sexto de primaria (2008) y tercero de secundaria (2011), a partir de datos de la prueba ENLACE aplicada a la población de estudiantes del Estado de Tlaxcala; comparar el grado de logro académico en español y matemáticas en 2008 y 2011; y evaluar la importancia relativa de 13 variables predictoras: puntaje en matemáticas, puntaje en español, sexo, beca, turno de la escuela, tipo de escuela, marginación, ubicación geográfica (altitud, latitud, longitud).

## 1.4 MATERIALES Y MÉTODOS

### 1.4.1 Colecta y preparación de datos

La base de datos de registros académicos utilizados en este estudio (2008-2011) para el Estado de Tlaxcala corresponde a un subconjunto de la base datos nacional de la prueba ENLACE que se aplicó de 2006 a 2014 a todos los alumnos de 3° de primaria a 3° de bachillerato cuyo propósito fue generar información para los padres de familia, docentes, directivos y a la sociedad en general sobre el logro académico de los estudiantes del Sistema Educativo, en español y matemáticas (SEP, 2008). Los registros se almacenaron por año de aplicación, nivel educativo y grado, en donde cada registro pertenece a un alumno; también se encuentran variables, que informan del logro académico de los alumnos que depende de los cortes en sus puntajes (Cuadro 1.1), el puntaje obtenido en la prueba (en una escala de 200 a 800), si el alumno fue becario, turno de la escuela, tipo de sostenimiento de la escuela, y variables de localización de la escuela.

**Cuadro 1.1. Intervalos de puntaje para determinar el nivel de logro académico en español y matemáticas.**

Grado y año	Español		
	0 <sup>†</sup>	1 <sup>¶</sup>	2 <sup>§</sup>
6° 2008	(413.85]	(413.85, 581.62]	(581.63, 714.01)
7° <sup>p</sup> 2009	(446.31]	(446.31, 593.18]	(593.18, 735.69)
8° <sup>m</sup> 2010	(445.08]	(445.08, 592.41]	(592.41, 735.35)
9° <sup>††</sup> 2011	(462.94]	(462.94, 608.22]	(608.22, 749.18)
	Matemáticas		
6° 2008	(412.62]	(412.62, 608.13]	(608.14, 735.70)
7° <sup>p</sup> 2009	(507.27]	(507.27, 634.85]	(634.85, 737.26)
8° <sup>m</sup> 2010	(505.39]	(505.39, 634.05]	(364.05, 737.32)
9° <sup>††</sup> 2011	(525.99]	(525.99, 657.03]	(657.03, 762.21)

†0: insuficiente; ¶1: elemental; §2: bueno o excelente.

<sup>p</sup>7°: 1° de secundaria, <sup>m</sup>8°: 2° de secundaria, 9°: <sup>††</sup>3° de secundaria.

**Fuente: Prueba ENLACE 2008-2013 (SEP, 2008).**

De la información disponible para el estado de Tlaxcala se seleccionó el subconjunto de alumnos que realizaron la prueba ENLACE durante cuatro años consecutivos (2008 a 2011), de 6° de primaria hasta 3° de secundaria, que marcan el inicio de una trayectoria educativa (el paso de primaria hasta llegar a 3° de secundaria). Las variables de contexto seleccionadas son: puntaje en español, puntaje en matemáticas, becario, turno de la escuela, tipo de sostenimiento, tipo de localidad, sexo, tipo de escuela, tamaño de localidad medida por sus habitantes, nivel de marginación y la localización geográfica (latitud, longitud y altitud) del municipio en donde se encuentra la escuela (Cuadro 1.2).

**Cuadro 1.2. Variables contextuales y localización geográfica seleccionadas del Estado de Tlaxcala, México.**

Variable	Descripción	Valores	Fuente de información
<i>n_esp</i>	Logro académico en español	0: insuficiente; elemental; 1: bueno o excelente	SEP, 2008
<i>n_mat</i>	Logro académico en matemáticas	0: insuficiente; elemental; 1: bueno o excelente	
<i>p_esp</i>	Puntaje en español	Rango entre 200 a 800 puntos	
<i>p_mat</i>	Puntaje en matemáticas	Rango entre 200 a 800 puntos	
<i>becario</i>	Condición de becario	0: no becario, 1: becario	
<i>turno</i>	Turno de la escuela	0: matutino, 1: vespertino	
<i>t_sost</i>	Tipo de sostenimiento	0: público, 1: privado	
<i>t_loc</i>	Tipo de la localidad	0: urbano, 1: rural	
<i>sexo</i>	Sexo	0: hombre, 1: mujer	
<i>t_esc</i>	Tipo de escuela	1: general, 2: indígena, 3: conafe, 4: particular, 5: telesecundaria, 6: técnica	
<i>t_loc</i>	Tamaño de la localidad	1: menos de 100 hab, 2: 100 hab a 249 hab, 3: 250 hab a 499 hab, 4: 500 hab a 2499 hab, 5: 2500 hab a 14 999 hab, 6: 15 000 y más hab.	
<i>n_mar</i>	Nivel de marginación	1: muy alto, 2: alto, 3: medio, 4: bajo, 5: muy bajo	
<i>lat</i>	Latitud	Decimal	INEGI, 2021

Variable	Descripción	Valores	Fuente de información
<i>lon</i>	Longitud	Decimal	
<i>alt</i>	Altitud	Metros sobre el nivel del mar	
<i>pob</i>	Población	En miles de habitantes	

**Fuente:** Elaboración propia con base en los formatos F911 de la SEP, información de ENLACE y con datos del Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Localidades del INEGI.

En el Cuadro 1.3 se presenta la descripción de los cuatro conjuntos de datos analizados por asignatura y año para la misma población escolar. Con el propósito de reducir el desbalance entre clases o niveles de logro académico los niveles bueno y excelente se agruparon en el nivel 2 de logro académico.

**Cuadro 1.3. Descripción de los conjuntos datos analizados por asignatura, año y niveles de logro académico en el estado de Tlaxcala para un total de 11 036 registros de estudiantes.**

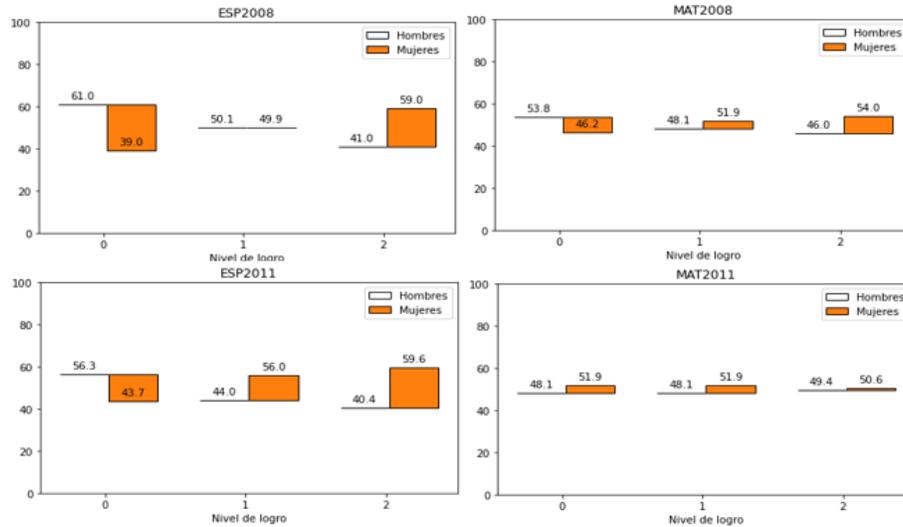
Año y asignatura evaluada	Niveles de logro académico		
	0 <sup>†</sup>	1 <sup>¶</sup>	2 <sup>§</sup>
ESP2008 <sup>p</sup>	1274	5949	3813
ESP2011 <sup>q</sup>	4194	5362	1480
MAT2008 <sup>††</sup>	1544	6039	3453
MAT2011 <sup>¶¶</sup>	6254	3608	1174

<sup>†</sup>0: insuficiente; <sup>¶</sup>1: elemental; <sup>§</sup>2: bueno o excelente.

<sup>p</sup> Español 2008, <sup>q</sup> Español 2011, <sup>††</sup> Matemáticas 2008, <sup>¶¶</sup> Matemáticas 2011.

**Fuente:** Elaboración propia con base en los datos en ENLACE 2008 y 2011 (SEP, 2008).

En la Figura 1.1 se muestra la distribución de los niveles de logro por sexo. En ESP2008 hay una proporción mayor de mujeres en la clase 2 y MAT2008 sobresalen también en las clases 1 y 2. Lo mismo se observa en ESP2011 y en MAT2011 en donde las mujeres sobresalen en ambas clases.



**Figura 1.1. Niveles de logro académico por sexo para los cuatro conjuntos de datos, ESP2008, ESP2011, MAT2008 y MAT2011. Fuente: elaboración propia.**

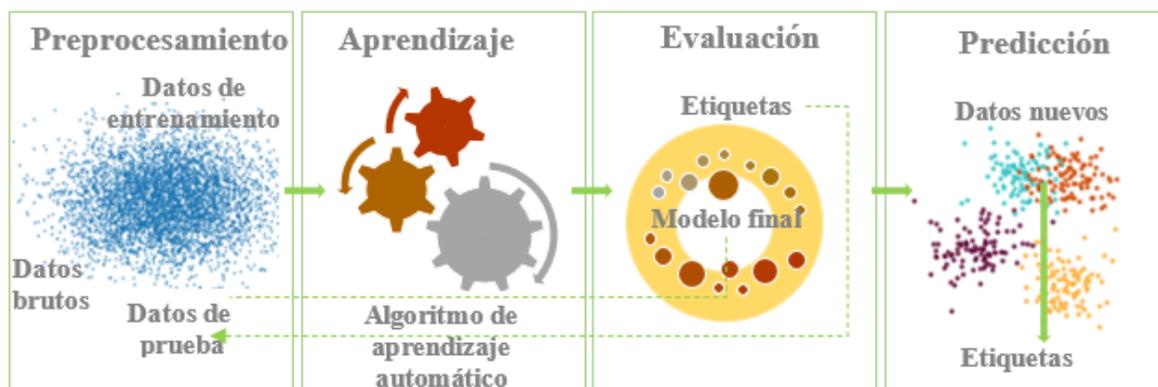
De la información preliminar de la prueba ENLACE para Tlaxcala se eliminaron 16 registros que tienen la misma información en todas las variables. Las variables predictoras categóricas  $t\_esc$ ,  $t\_loc$  y  $n\_mar$ , se transformaron para su análisis en variables indicadoras. Los cuatro conjuntos de datos depurados consisten de 22 variables predictoras y 11 036 registros de alumnos que representa 44.3% del total de 24 875 registros de alumnos evaluados en 2008 en el Estado de Tlaxcala. Para el análisis se generaron cuatro conjuntos de datos: ESP2008, MAT2008, ESP2011 y MAT2011. En cada conjunto de datos las etiquetas de clase corresponden al nivel de logro de los alumnos obtenido en la materia correspondiente,  $n\_esp$  o  $n\_mat$ .

Para el procesamiento y creación de los cuatro conjuntos de datos se usó el software (SAS, Statistical Analysis System) v. 9.4. Para correr los algoritmos de aprendizaje automático se empleó el software Python 3.8 a través de la biblioteca Scikit-learn para la ejecución de los códigos y la generación de resultados.

### 1.4.2 Clasificadores de aprendizaje automático

El objetivo principal de los modelos de aprendizaje automático supervisado es generar un modelo predictivo con base en datos de entrada y clases objetivo, definidas previamente. Es decir, se define un conjunto de muestras asociadas con su clase objetivo y con este conjunto de datos de entrenamiento se realiza el ajuste inicial del modelo. De esta forma el algoritmo

aprende a clasificar las muestras de entrada, el modelo compara la clase predicha con la clase objetivo de la muestra para realizar el ajuste de los pesos del modelo, de acuerdo con los errores de clasificación.

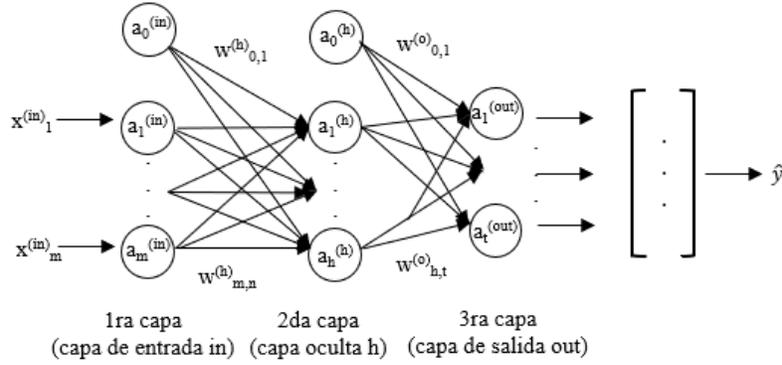


**Figura 1.2. Proceso general del aprendizaje automático supervisado. Fuente: elaboración propia con base en Raschka y Mirjalili 2017.**

En este trabajo se implementaron tres clasificadores de aprendizaje supervisado, una red neuronal multicapa (MLP, multilayer perceptron) y un algoritmo de ensamble potenciación del gradiente (GB, gradient boosting) para predecir el nivel del logro académico y un modelo bosque aleatorio (RF, random forest) para determinar la importancia relativa de las variables predictoras. El objetivo es que el modelo aprenda a partir de datos de entrenamiento que están asociados a una clase. El modelo una vez entrenado es capaz de predecir la clase correspondiente a un nuevo conjunto de datos de entrada (Raschka y Mirjalili, 2017).

#### 1.4.2.1 Red neuronal multicapa

El clasificador de red neuronal multicapa, es una red de neuronas conectadas por pesos o parámetros, estructuradas en capas de entrada (in), capas ocultas (h) y capas de salida (out), la arquitectura básica de MLP es de tres capas (Figura 1.3). Entre más capas tenga la red y más compleja sea, también serán más complejas las funciones que pueda realizar (Borkar y Rajeswari, 2014).



**Figura 1.3. Arquitectura de un clasificador de red neuronal multicapa (MLP). Fuente: Raschka y Mirjalili, 2017.**

El funcionamiento de MLP consiste en que, dado un vector de datos de entrada  $x_i$  con  $m$  variables predictoras, en la capa oculta se infieren  $M$  funciones; luego, en la capa de salida se determina la respuesta predicha, al aplicar las funciones inferidas en la capa oculta por medio de una transformación no lineal (González-Camacho *et al.*, 2012). Todas las capas están conectadas hacia adelante, y se representan por una matriz de pesos  $W$ , que se inicializa con valores aleatorios pequeños. Primero se activa la capa oculta  $a_1^{(h)}$  por medio de una función de activación  $\phi$  (por ejemplo, una función sigmoide) a los valores de  $Z$  (matriz de valores netos) que resulta de la combinación lineal de las variables de entrada con los pesos  $W$ ; estos valores son las entradas de la capa de salida. La activación de la capa oculta se realiza con las siguientes expresiones (Raschka y Mirjalili, 2017):

$$Z^{(h)} = A^{(in)}W^{(h)}, \quad A^{(h)} = \phi(Z^{(h)})$$

donde  $A^{(in)}$  es una matriz de las características de las muestras  $x^{(in)}$ ;  $W^{(h)}$  es la matriz de pesos y  $\phi(\cdot)$  es la función de activación. De manera similar, se genera la activación de la capa de salida:

$$Z^{(out)} = A^{(h)}W^{(out)}, \quad A^{(out)} = \phi(Z^{(out)})$$

donde  $W^{(out)}$  es una matriz de pesos de salida y  $A^{(out)}$  es una matriz de probabilidades con las respuestas o clases predichas de la red. Para determinar el error de clasificación, se compara la salida deseada con la predicha. Se utiliza el algoritmo de propagación hacia atrás para distribuir los errores y se obtienen las derivadas parciales con respecto a los pesos de la red para actualizar el modelo (Raschka and Mirjalili, 2017).

### 1.4.2.2 Clasificador potenciación del gradiente

El clasificador potenciación del gradiente consiste en un conjunto de árboles de decisión individuales que se entrenan de manera secuencial, tal que cada árbol mejore los errores de los árboles anteriores. Para predecir una nueva observación se agregan las predicciones de todos los árboles individuales del modelo. GB puede emplear cualquier función de pérdida siempre que ésta sea diferenciable. Se ajusta un modelo, por ejemplo,  $f_1$  para predecir la variable respuesta, después se calculan los errores  $y - f_1(x)$ ; luego, se ajusta un modelo  $f_2$  que intenta predecir los errores del modelo anterior; nuevamente se ajusta un modelo  $f_3$  que trata de corregir los errores de los modelos anteriores y esto se repite  $m$  veces. Para evitar sobre ajuste se emplea un parámetro de regularización que se denomina tasa de aprendizaje ( $\lambda$ ) que limita la influencia de cada modelo en el conjunto de ensamble.

$$y \approx \lambda \sum_{i=1}^m f_i(x)$$

La idea detrás de la potenciación es ajustar de forma secuencial múltiples modelos sencillos, en que cada modelo emplea información del modelo anterior para “aprender de sus errores” y mejorar en cada iteración y como valor final se toma el promedio de las predicciones (Rogers y Gunn, 2005).

### 1.4.2.3 Bosque aleatorio

Los bosques aleatorios (RF) son un conjunto de árboles de decisión y su objetivo es promediar los múltiples arboles de decisión para construir un modelo más robusto que tenga una mejor generalización y sea menos susceptible al sobre ajuste (Raschka y Mirjalili, 2017). Para predecir la clase se usan las reglas de cada árbol y se asignan por mayoría de votos. (Breiman, 2001). El algoritmo de RF se resume como sigue:

Se selecciona una muestra de tamaño  $n$  del conjunto de variables predictoras (mediante muestreo aleatorio sin reemplazo, bootstrap), el árbol crece a partir de una muestra inicial; para cada nodo se seleccionan  $d$  características sin reemplazo; se divide el nodo con la función que proporcione la mejor división de acuerdo con la función ganancia de información ( $IG$ , information gain) que se define como:

$$IG = (D_p, f) = I(D_p) - \frac{N_l}{N_p} I(D_l) - \frac{N_r}{N_p} I(D_r)$$

donde  $f$  es la característica para realizar la división;  $N_p$  es el número total de muestras en el nodo padre;  $N_l$  es el número de muestras en el nodo hijo izquierdo;  $N_r$  es el número de muestras en el nodo hijo derecho;  $I$  es la medida de impureza (gini, entropía o error de clasificación);  $D_p$  es el conjunto de datos en el nodo padre;  $D_l$  es el conjunto de datos en el nodo hijo izquierdo;  $D_r$  es el conjunto de datos en el nodo hijo derecho.

Los bosques aleatorios también se usan para determinar la importancia de un conjunto de variables en el modelo. El algoritmo de RF crea clasificadores con una selección aleatoria de las características; esto logra una buena exploración de subconjuntos de éstas; en donde se seleccionan aquellas con mayor importancia (Rogers y Gunn, 2005).

### 1.4.3 Criterios de desempeño de los modelos de predicción

Para evaluar el desempeño de los modelos de clasificación MLP, GB y RF las métricas se obtienen básicamente a partir de una matriz de confusión ( $MC$ ). La cual, describe el conteo de los verdaderos positivos ( $VP$ ), verdaderos negativos ( $VN$ ), falsos positivos ( $FP$ ) y falsos negativos ( $FN$ ). Las filas representan número de muestras en la clase observada y las columnas el número de predicciones de cada clase. La diagonal principal de  $MC$  corresponde al número de muestras que el algoritmo clasifica correctamente en cada clase. Si  $MC$  solo tiene valores positivos en la diagonal, indica que el clasificador clasifica correctamente todas las muestras. La métrica precisión global de clasificación ( $PG$ ) mide la proporción de muestras bien clasificadas en cada clase y se calcula como:

$$PG = \frac{VP + VN}{FP + FN + VP + VN}$$

Las métricas para medir el desempeño del clasificador en cada clase son precisión ( $P$ ), sensibilidad ( $S$ ), especificidad ( $E$ ) y  $FI$ . Se definen con las siguientes expresiones:

$$P = \frac{VP}{VP + FP}$$

$$S = \frac{VP}{FN + VP}$$

$$E = \frac{VN}{VN + FP}$$

$$F1 = 2 \frac{P \times S}{P + S}$$

En donde el valor de *F1* resume la *P* y *S* en una sola métrica y es de gran utilidad en clases desbalanceadas y varía entre 0 y 1. La curva ROC (Receiver Operating Characteristics) es una curva que grafica la *S* versus *1-E*. Los diferentes puntos en la curva corresponden a los puntos de corte utilizados para determinar si los resultados de la prueba son positivos.  $AUC_{ROC}$  (área bajo la curva ROC) su valor se interpreta como la probabilidad de que ante dos muestras una positiva y una negativa, la prueba asigne una probabilidad más alta a la muestra positiva, clasificación correcta (Mandrekar, 2010). Su valor oscila entre 0 y 1, entre mayor sea el  $AUC_{ROC}$  mejor será la clasificación, un valor cercano a 0.50 indica una mala clasificación. La curva *P-S* es el resultado de graficar la *P* y *S*. Ésta permite ver a partir de que *S* se tiene una degradación de la *P* y viceversa. El resultado ideal es una curva que se acerque a la esquina superior derecha (alta *P* y *S*), lo que genera un área bajo la curva  $AUC_{P-S}$  que entre más cercano a 1 sea, mejor será el modelo (Saito y Rehmsmeier, 2015).

#### 1.4.4 Entrenamiento y validación de los modelos

Para hacer el entrenamiento de los modelos, cada conjunto de datos se dividió en dos, 80.0% para entrenamiento y 20.0% prueba. Para cada conjunto de datos (ESP2008, MAT2008, ESP2011 y MAT2011) se aplicó el clasificador RF para evaluar la importancia relativa de las características de entrada. En la prueba de los modelos se consideró 20.0% de los datos y se presenta el caso de clases desbalanceadas en los extremos (Cuadro 1.4).

**Cuadro 1.4. Número de muestras del conjunto de prueba por clase objetivo.**

Clase	ESP2008	MAT2008	ESP2011	MAT2011
0 <sup>†</sup>	255	309	839	1251
1 <sup>¶</sup>	1190	1208	1073	722
2 <sup>§</sup>	763	691	296	235

<sup>†</sup>0: insuficiente; <sup>¶</sup>1: elemental; <sup>§</sup>2: bueno o excelente.

**Fuente: Elaboración propia con base en los datos de ENLACE 2008-2011.**

##### 1.4.4.1 Selección de hiperparámetros óptimos

En los modelos de aprendizaje automático, los hiperparámetros son parámetros que pueden ser ajustados para controlar el proceso de entrenamiento del modelo. El ajuste de

hiperparámetros consiste en el proceso de encontrar la combinación de hiperparámetros, de un conjunto de valores, que generen el mejor desempeño de clasificación, en este caso a través de la mejor *PG*. Con 80.0% de los datos se realiza el ajuste de hiperparámetros utilizando validación cruzada (CV, cross-validation) con cinco particiones aleatorias y estratificadas por clase objetivo. El método de validación cruzada consiste en subdividir el conjunto de entrenamiento (80.0% del conjunto total) en  $k$  subconjuntos del mismo tamaño. Se itera a través de los subconjuntos y en cada iteración se usa uno de los subconjuntos como conjunto de validación y los restantes como conjunto de entrenamiento. Este proceso se hace hasta que cada subconjunto se haya usado como muestra de prueba. Se generaron cinco grupos disjuntos cada uno con 20.0% de los datos; y en la primera iteración se toman cuatro grupos de entrenamiento (80.0%) y uno de prueba (20.0%). Para ajustar los hiperparámetros de los dos modelos se usaron distintos valores que se muestran en el Cuadro 1.5.

**Cuadro 1.5. Intervalo de valores para la prueba de hiperparámetros.**

Modelo	Hiperparámetros	Intervalos de valores
MLP	<i>nco</i>	50, 100, 200, 250
	<i>fa</i>	Tanh, relu, logistic
	<i>op</i>	Sgd, Adam, lbfgs
	<i>re</i>	0.0001, 0.001, 0.1
	<i>ta</i>	Constant, adaptative
GB	<i>mi</i>	90, 100, 110
	<i>pa</i>	4, 5, 6
	<i>ha</i>	3, 4

*nco*: neuronas en capa oculta; *fa*: función de activación; *op*: optimizador de pesos; *re*: regularizador; *ta*: tasa de aprendizaje; *mi*: máximo de iteraciones; *pa*: profundidad de árbol; *ha*: hojas por árbol. **Fuente: Elaboración propia.**

#### 1.4.4.2 Evaluación del desempeño de los modelos de clasificación

Una vez que se obtienen los valores óptimos de los hiperparámetros, se hace la evaluación de desempeño de los modelos para cada conjunto de datos. Para estimar *PG* de cada modelo se consideraron los datos completos (Cuadro 1.3) y se utiliza validación cruzada con  $k = 5$  particiones, es decir, el conjunto de datos de entrada se divide en cinco, de estos un grupo se

utilizó para prueba y los cuatro restantes para entrenamiento (80.0% para entrenamiento y 20.0% para prueba en predicción). En cada iteración se obtuvo un valor  $PG$  del modelo; al finalizar las iteraciones, se hace un promedio de los valores de  $PG$  para tener un valor promedio de  $PG$  y su desviación estándar y las demás métricas propuestas en el estudio.

#### **1.4.5 Importancia relativa de las características de entrada**

Existen algoritmos para determinar la relevancia de diferentes características en términos de cuan útiles pueden ser para predecir la salida deseada u objetivo. El algoritmo de RF construye una gran cantidad de clasificadores con características que se seleccionan al azar logrando con esto explorar posibles subconjuntos de características. En cada nodo de RF se selecciona una característica y se usa para dividir el nodo y maximizar la ganancia de información. Las medidas de importancia de variables se pueden usar para el rendimiento del algoritmo de aprendizaje automático (Rogers y Gunn, 2005).

La determinación de la importancia relativa de las 22 variables predictoras se hizo con RF. Mediante validación cruzada (CV) con cinco particiones se determinaron los valores óptimos de los hiperparámetros del modelo. Posteriormente se ajusta el modelo para determinar la importancia de las variables a través de la biblioteca Scikit-learn de Python con feature importance. Este procedimiento se realizó para los cuatro conjuntos de datos: ESP2008, MAT2008, ESP2011 y MAT2011.

## 1.5 RESULTADOS Y DISCUSIÓN

La generación de los resultados de esta sección, se elaboraron con los scripts en Python descritos en el Anexo.

### 1.5.1 Hiperparámetros óptimos

Los hiperparámetros óptimos de los modelos se obtuvieron mediante validación cruzada y búsqueda en retícula. En el Cuadro 1.6 se muestran los resultados obtenidos para cada grupo de datos.

**Cuadro 1.6. Hiperparámetros óptimos para el clasificador red neuronal multicapa y potenciación del gradiente y cada conjunto de datos (asignatura y año).**

Modelo	Hiperparámetros	ESP2008	ESP2011	MAT2008	MAT2011
MLP	<i>nco</i>	100	200	50	30
	<i>fa</i>	Logistic	Tanh	Relu	Logistic
	<i>op</i>	Adam	Sgd	Sgd	Adam
	<i>re</i>	0.1	0.1	0.1	0.0001
	<i>ta</i>	Constant	Adaptative	Constant	Constant
GB	<i>mi</i>	90	90	90	110
	<i>pa</i>	4	4	4	4
	<i>ha</i>	3	4	3	4

*nco*: neuronas en capa oculta; *fa*: función de activación; *op*: optimizador de pesos; *re*: regularizador; *ta*: tasa de aprendizaje; *mi*: máximo de iteraciones; *pa*: profundidad de árbol; *ha*: hojas por árbol. **Fuente: Elaboración propia.**

### 1.5.2 Desempeño de los clasificadores

Sobre los resultados de los modelos, se presentan en cuadros comparativas por materia evaluada y año de la prueba ENLACE. Para cada modelo se presenta la *PG* del modelo con su desviación estándar, la curva ROC y las curvas de *P-S*.

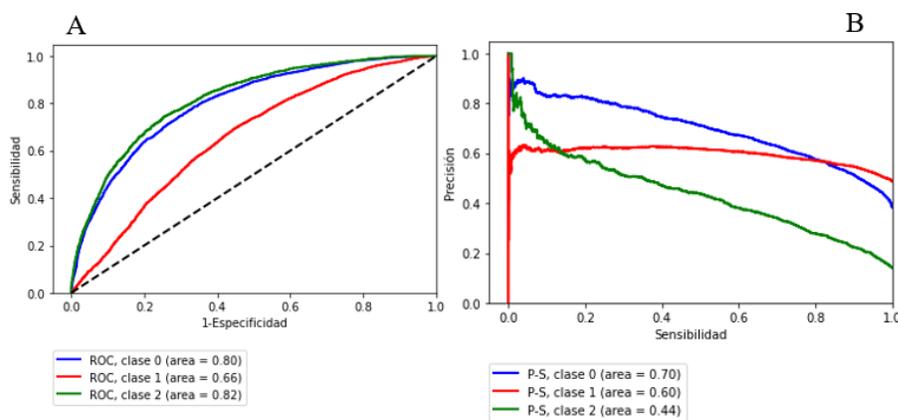
**Cuadro 1.7. Precisión global de clasificación de los modelos por año y materia evaluada.**

Modelo	ESP2008	ESP2011	MAT2008	MAT2011
MLP	0.701 (+- 0.033)	0.611 (+- 0.019)	0.672 (+- 0.034)	0.631 (+- 0.008)
GB	0.695 (+- 0.024)	0.610 (+- 0.016)	0.688 (+- 0.008)	0.635 (+- 0.006)

**Fuente: Elaboración propia.**

De los dos modelos que se aplicaron a los datos, el que mejor predice la clasificación de los alumnos, en español 2008, en cada una de las clases, es MLP con  $PG$  de 70.1% (de los datos se clasifican correctamente en la clase en que la SEP los clasificó originalmente) (Cuadro 1.7). De la  $MC$  para el mejor modelo, la clase 1 (elemental) y la clase 2 (bueno o excelente) son las clases que mejor clasifican a los alumnos, en tanto la mayoría de las muestras de las clases 0 se clasifican como muestras de clase 1. Con la información de  $P$  por clases se tiene que, efectivamente las clases 1 y 2 con las que resultan mejor clasificadas con valores de 71.0% y 80.0% respectivamente. Lo anterior se refuerza al observar los resultados de  $AUC_{P-S}$  en la que lo ideal es que las curvas por clase se acerquen lo máximo posible a la esquina superior derecha (se acerca a 1), es decir una alta  $P$  y  $S$ . Se observan valores de 0.72 y 0.75 para las clases 1 y 2 (Cuadro 1.8). Considerando la métrica de  $FI$ , se tiene un valor de 0.77 y 0.75 para las clases 1 y 2 lo que es aceptable y significa que el algoritmo MLP detecta bien las clases, aunque también incluye muestras de las otras clases.

En español 2011, los dos modelos propuestos clasificaron con una  $PG$  similar (61.0% de los datos clasificados correctamente en las clases que les corresponden). MLP tiene  $PG$  de 61.1% y GB de 61.0%. Para MLP, la clase 0 y la clase 1 son las que mejor clasifican las muestras con  $P$  por clase de 67.0% y 61.0% (más de 60.0% de las muestras se clasifican correctamente). Para este año más de 50.0% de las muestras de la clase 2 se clasificaron como clase 1 y considerando el valor de  $FI$  para esta clase que es de 0.41 se tiene que el modelo no detecta la clase muy bien. Lo mismo se observa en la gráfica de las curvas  $P-S$  con un valor bajo para la clase 2 (Figura 1.4).



**Figura 1.4. Curvas para el modelo MLP-español 2011. A) B) curva ROC, B) curva de P-S.**

**Fuente: Elaboración propia.**

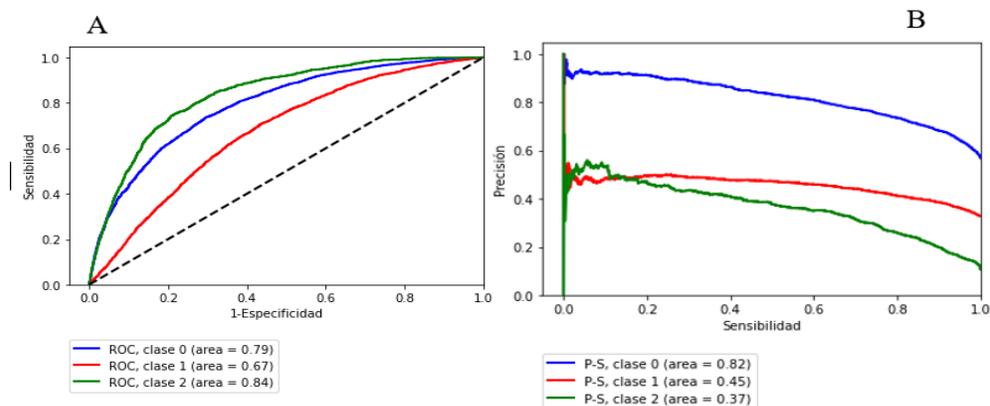
**Cuadro 1.8. Valores de P y F1, AUC<sub>ROC</sub> y AUC<sub>P-S</sub> para cada modelo por año y clase en español.**

Modelo	Clase	ESP2008			ESP2011		
		<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>	<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>
MLP	0	0.32	0.86	0.46	0.65	0.80	0.70
	1	0.77	0.74	0.72	0.67	0.66	0.60
	2	0.75	0.87	0.75	0.41	0.82	0.44
GB	0	0.33	0.86	0.40	0.65	0.80	0.69
	1	0.77	0.71	0.68	0.67	0.64	0.57
	2	0.75	0.87	0.77	0.42	0.81	0.41

*P*: precisión; *FI*: f1-score; AUC<sub>ROC</sub>: Área bajo la curva ROC; AUC<sub>P-S</sub>: Área bajo la curva *P-S*. **Fuente:** Elaboración propia.

Para las muestras de matemáticas 2008, el modelo que mejor predice la clasificación es GB con una precisión de 68.8% (los datos se clasifican correctamente en la clase en que la SEP los clasificó originalmente). Considerando las 22 variables seleccionadas que se tomaron como entrada de los modelos, MLP tiene *PG* de 67.2% y GB de 68.8%. Las clases 1 y 2 con las que resultan mejor clasificadas con *P* de 70.0% y 79.0%, seguidas de la clase 0 con 67.0% de muestras correctamente clasificadas. De la gráfica de *P-S* se observan valores AUC<sub>P-S</sub> por arriba de 0.68 para las clases 1 y 2 y 0.50 para la clase 0. Los valores de *FI* muestran que GB detecta bien las clases 1 y 2 aunque incluyen muestras de las otras clases.

En matemáticas 2011, GB fue el mejor clasificador con *PG* de 63.5% y una desviación estándar de 0.006. La clase 0 y la clase 1 son las que se clasificaron mejor ya que presentan una *P* de 72.0% y 49.0%. De los resultados de AUC<sub>P-S</sub> se obtiene un valor bajo para la clase 2 que fue la que peor clasificó, 0.37 (Figura 1.5).



**Figura 1.5.** Ejemplo de curvas para el modelo GB-matemáticas 2011. A) curva ROC B) curva P-S. Fuente: Elaboración propia.

**Cuadro 1.9.** Valores de P y F1,  $AUC_{ROC}$  y  $AUC_{P-S}$  para cada modelo por año y clase en matemáticas.

Modelo	Clase	<i>FI</i>	MAT2008		MAT2011		
			$AUC_{ROC}$	$AUC_{P-S}$	<i>FI</i>	$AUC_{ROC}$	$AUC_{P-S}$
MLP	0	0.43	0.86	0.50	0.76	0.79	0.82
	1	0.77	0.70	0.68	0.44	0.68	0.47
	2	0.74	0.86	0.73	0.30	0.83	0.38
GB	0	0.39	0.86	0.48	0.77	0.79	0.82
	1	0.78	0.70	0.69	0.46	0.67	0.45
	2	0.73	0.87	0.76	0.39	0.84	0.37

*P*: precisión; *FI*: f1-score;  $AUC_{ROC}$ : Área bajo la curva ROC;  $AUC_{P-S}$ : Área bajo la curva P-S. Fuente: Elaboración propia.

### 1.5.3 Importancia relativa de características

Al aplicar RF a los cuatro grupos de datos se encontró que, en español y matemáticas, 2008 y 2011, las variables de calificación  $p\_esp$  y  $p\_mat$ , son las más importantes (Figura 1.6, Figura 1.7), sin embargo, evaluando los dos modelos seleccionados y considerado estas variables y siete más (*becario*, *turno*, *sexo*, *lat*, *lon*, *alt*, *pob*) la PG de los modelos de clasificación no mejoró, y se decidió considerar todo el conjunto de variables y la partición propuesta de 80.0% y 20.0%.

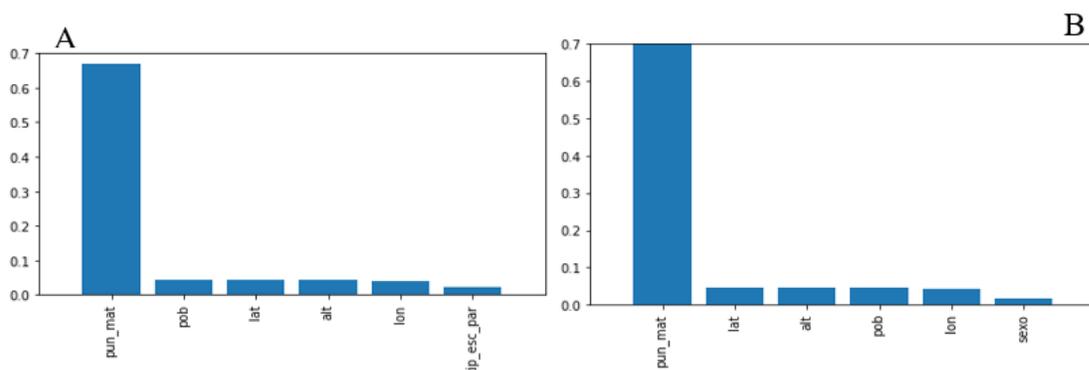


Figura 1.6. Importancia de seis variables con RF. A) Español 2008, B) Matemáticas 2008, fuente: Elaboración propia.

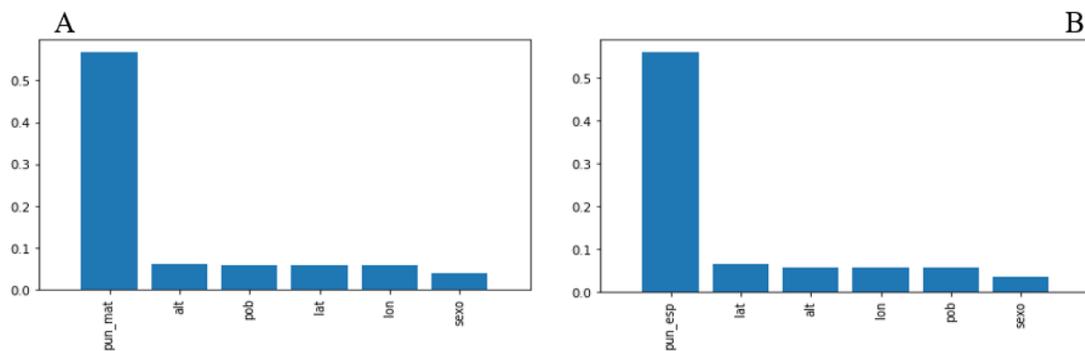


Figura 1.7. Importancia de seis variables con RF. A) Español 2011, B) Matemáticas 2011, fuente: Elaboración propia.

Además de las variables que corresponden a los puntajes en la prueba, las variables de *sexo*, *t\_esc particular*, *lat*, *lon*, *alt* y *pob* resultan importantes dentro de los mediante RF.

### 1.5.4 Discusión

De los resultados se tiene que con las 22 variables que se manejaron para hacer la clasificación y predecir los niveles de logro de los estudiantes, sí se observa una relación de éstas con la clase que se asigna al desempeño obtenido. Una de las variables más importantes

fue usar el puntaje obtenido de matemáticas para predecir la clasificación en español y usar los puntajes de español en la clasificación de matemáticas. En orden de importancia, aunque en menor medida, la ubicación geográfica de la escuela, la población de la localidad de ubicación de la escuela y el sexo del alumno. Se destaca la influencia que se observó de las variables  $p\_esp$  y  $p\_mat$  en los resultados obtenidos para la clasificación de los alumnos en los niveles de logro. En estudios revisados se señala al español como un aprendizaje que resulta de un proceso de acumulación de experiencias pedagógicas que el alumno tiene en su estancia en la escuela y al aprendizaje de matemáticas como un proceso constructivo que se relaciona con la formulación o comprensión de conceptos con la resolución de problemas. Es importante observar las calificaciones de los estudiantes en español para poder predecir la clasificación de los niveles de logro en matemáticas y viceversa, es decir, conviene saber qué pasó en la otra materia como una medida general de capacidad de los alumnos.

Con las diferentes métricas seleccionadas para comparar los dos modelos, el área bajo la curva  $AUC_{P-S}$  dieron mejor información para evaluar el rendimiento de los clasificadores al discriminar el porcentaje de muestras clasificadas correctamente y los resultados son coincidentes respecto a la matriz de confusión.

Al aplicar los dos algoritmos de aprendizaje automático supervisado (MLP y GB), en general se observó una buena clasificación para ambos modelos ( $PG$  mayor al 60.0%) en todas las clases, lo que se constata con los resultados de las métricas seleccionadas para evaluar ambos modelos de manera global. Sin embargo, en las limitaciones del trabajo se encuentra el hecho de tener clases desbalanceadas ya que esto influye en que algunas clases incluyan muestras de las otras clases en la predicción. Para mejorar el trabajo se pueden considerar variables de contexto adicionales y observar si éstas mejoran la clasificación.

## 1.6 CONCLUSIONES

Con base en los objetivos y resultados, de los dos modelos empleados para clasificación, en español 2008 y 2011, el clasificador MLP alcanzó una precisión global de clasificación correcta (*PG*) de 61.0% de las muestras correctamente clasificadas. En matemáticas fue el clasificador potenciación del gradiente (*GB*) que clasificó, en 2008, el 68.8% de las muestras correctamente y 63.5% en 2011. El desempeño de los clasificadores por clase objetivo fue superior para las clases o grados de logro académico 1 y 2. En general se observó una *PG* de los dos modelos aplicados de 60.0% pero es necesario identificar si es pertinente incluir variables adicionales para aumentar esta *PG*.

El conjunto de variables contextuales utilizadas en este estudio muestra una asociación con los grados de logro académico de los estudiantes (asociación del nivel de logro). En particular respecto a *becario*, el *sexo* y el *turno* de la escuela. Además, se observó que el puntaje en español de un estudiante condiciona fuertemente su grado de logro académico en matemáticas. Esto es importante ya que su identificación puede ser un indicio para la generación de políticas públicas orientadas a la atención de aquellos alumnos que se identifiquen con un mal desempeño y poder darles un seguimiento y evitar su salida del sistema educativo.

## **CAPÍTULO II. COMPARAR LA PREDICCIÓN DE LOS NIVELES DE LOGRO ACADÉMICO EN EL NIVEL BÁSICO EN LOS ESTADOS DE TLAXCALA, GUERRERO Y NUEVO LEÓN CON BASE EN ALGORITMOS DE APRENDIZAJE AUTOMÁTICO**

### **2.1 RESUMEN**

Uno de los intereses más importantes de las escuelas es que los alumnos terminen con éxito su educación, sin embargo, hay factores dentro y fuera de la escuela que condicionan o limitan un tránsito educativo ideal. En este trabajo se implementaron dos algoritmos de aprendizaje automático supervisado, una red neuronal multicapa (MLP) y un modelo de potenciación del gradiente (GB) para predecir el rendimiento escolar en español y matemáticas (2008-2011) de alumnos de educación básica y contrastar el nivel de logro académico en los estados de Tlaxcala, Guerrero y Nuevo León con base en variables contextuales obtenidas de la prueba ENLACE. Las variables fueron: puntaje en español, puntaje en matemáticas, turno escolar, tipo de escuela, sexo, becario, entre otras, para determinar su asociación a tres clases o categorías de logro académico (0: insuficiente; 1: elemental; 2: bueno o excelente) propuestos por la Secretaría de Educación Pública (SEP). Los clasificadores MLP y GB se entrenaron y probaron con una base de datos de 11 036 registros en Tlaxcala, 32 382 para Guerrero y 40 935 en Nuevo León. MLP y GB alcanzaron una *PG* de más de 65.0% en español y matemáticas de 2008 y de 60.0% para ambas materias evaluadas en los dos años, pero el modelo que mejor clasificó fue GB. Del conjunto de variables predictoras se observa una asociación de los puntajes de logro en español y matemáticas, también resultó importante la población y el sexo de los alumnos. En 2008 para las dos materias evaluadas, Tlaxcala fue la entidad que obtuvo las mayores *P*; 71.0% en la clase 1 y 80.0% para la clase 2 en español. En 2011 las clases 0 y 2 tuvieron las mayores *P*, y fue para matemáticas en Nuevo León, con *P* de 75.0% para la clase 0 y 63.0% para la clase 2. En las tres entidades se observa que, en español 2008 se obtienen mejores resultados con MLP. En matemáticas se obtienen las mejores *P* con GB.

Respecto a los niveles de logro observados en las entidades por sexo, las mujeres sobresalen en porcentaje en el nivel de logro 2 para español en todas las entidades. En

matemáticas para Tlaxcala y Guerrero resaltan las mujeres en los niveles de logro superiores, pero en Nuevo León son los hombres lo que tienen mayor proporción.

Poder determinar los resultados y determinar a qué clase pertenecen los alumnos, resulta importante como ayuda a las autoridades educativas en la identificación de estudiantes con bajos logros y que posiblemente requieren apoyo especial, para centrarse en ellos o enfatizar en la modificación de métodos de enseñanza.

**Palabras clave:** Árbol de decisión, clases, clasificación supervisada, logro académico, red neuronal.

## 2.2 ABSTRACT

One of the most important interests of schools is that students successfully complete their education, however, there are factors inside and outside the school that condition or limit an ideal educational journey. In this work, two supervised machine learning algorithms, a multilayer neural network (MLP) and a gradient boosting (GB) model were implemented to predict school performance in spanish and mathematics subjects (2008-2011) of elementary school students and contrast the level of academic achievement in the states of Tlaxcala, Guerrero and Nuevo León based on contextual variables obtained from the ENLACE test. The variables were: score in spanish, score in mathematics, school shift, type of school, sex, scholarship holder, among others, to determine its association with three classes or categories of academic achievement (0: insufficient; 1: elementary; 2: good or excellent) proposed by the Ministry of Public Education (SEP). The MLP and GB classifiers were trained and tested with a database of 11 036 for Tlaxcala, 32 382 for Guerrero and 40 935 for Nuevo León. MLP and GB achieved a *PG* of more than 65.0% in spanish and mathematics in 2008 and 60.0% for both subjects tested in the two years, but the best-ranked model was GB. From the set of predictor variables, an association of achievement scores in spanish and mathematics is observed, the population and gender of the students were also important. In 2008, for the two subjects evaluated, Tlaxcala was the entity that obtained the highest *P*: 71.0% in class 1 and 80.0% for class 2 in spanish. In 2011, classes 0 and 2 had the highest *P*, and it was for mathematics in Nuevo León, with *P* of 75.0% for class 0 and 63.0% for class 2. In the three entities it is observed that, in spanish, 2008 is obtained better results with MLP. In mathematics you get the best *P* with GB.

Regarding the levels of achievement observed in the entities, by sex, women stand out in percentage at achievement level 2 for spanish in all entities. In mathematics for Tlaxcala and Guerrero women stand out in the higher levels of achievement, but in Nuevo León it is men who have the highest proportion.

Being able to determine the results to which class the students belong is important as an aid to the educational authorities in identifying students with low achievements and who need special support, to focus on them or to emphasize the modification of teaching methods.

**Key words:** Decision tree, class, supervised classification, academic achievement, neural network.

## 2.3 INTRODUCCIÓN

En los años setenta toman relevancia las pruebas estandarizadas de rendimiento académico como herramientas para evaluar la calidad y el desarrollo educativo de un país; también son un medio para brindar información a tomadores de decisiones y encargados de la política educativa (Contreras Roldán y Backhoff, 2014). México ha participado recientemente en tres pruebas a gran escala, PISA, ENLACE y PLANEA; la primera que fue coordinada por la Organización para la Cooperación y el Desarrollo (OECD, Organisation for Economic Cooperation and Development) y las dos últimas desarrolladas por el propio país a través de la SEP, cuyos objetivos fueron informar a la sociedad y a las autoridades sobre el estado que guarda la educación nacional y brindar información oportuna a los docentes como apoyo a la mejora de su enseñanza, así como el aprendizaje de sus alumnos.

Diversos estudios relacionan la ubicación de la escuela, la escolaridad de los padres, y el número de libros en casa, por ejemplo, como factores que influyen en el nivel de desempeño alcanzado por los estudiantes en evaluaciones a gran escala. El nivel cultural familiar influye en las oportunidades educativas que se les brindan a los hijos (García-Jiménez *et al.*, 2020). En la actualidad se han desarrollado diversos estudios relacionados con la minería de datos y aprendizaje automático para proporcionar apoyo en los procesos de aprendizaje mediante el monitoreo de los estudiantes para medir su progreso (retención de alumnos y éxito de las escuelas). Polyzou *et al.*, (2018) identifican a dos grupos de estudiantes, aquellos con probabilidad de completar exitosamente un curso y los que tienen dificultades para finalizarlo; con esta identificación se puede proporcionar apoyo o recursos adicionales para incrementar su probabilidad de éxito. Seleccionan las calificaciones al final del semestre, las características específicas del estudiante y las características del curso. Para hacer la clasificación usan algoritmos de aprendizaje automático: árboles de decisión, máquinas de soporte vectorial, bosque aleatorio y potenciación del gradiente. Los métodos con mejores rendimientos fueron potenciación del gradiente y bosque aleatorio basados en las métricas  $AUC_{ROC}$  (área bajo la curva) y  $F1$ . Fakhri (2020) usa algoritmos de aprendizaje automático, bosques aleatorios, clasificador potenciación del gradiente, aprendizaje profundo, para predecir el rendimiento de los estudiantes (alto, medio, bajo) y seleccionan 17 características como edad, sexo, escolaridad de los padres, horas de estudio, etcétera. Se usó la matriz de

confusión,  $PG$  y  $AUC_{ROC}$  (área bajo la curva) para determinar el mejor clasificador. El modelo que mejor clasificó fue aprendizaje profundo con una  $PG$  de 0.782.

En México se han hecho diferentes estudios que tratan de determinar la relación de diversas variables internas y externas a la escuela con el desempeño de los estudiantes. Carvallo, *et al.*, (2007) analizan los efectos de diversas variables contextuales en una muestra de alumnos de sexto de primaria y tercero de secundaria de Baja California. A través de modelos jerárquicos lineales también se analizó el efecto que tiene el nivel socioeconómico, el contexto de la escuela y del profesor en el logro educativo de los estudiantes. Los resultados que obtuvieron fue que los alumnos del turno matutino, de escuelas privadas presentan mejor desempeño que sus pares del turno vespertino y de escuelas públicas. Quintana *et al.*, (2018) seleccionaron una muestra de alumnos del Centro Universitario UAEM Valle de México y usaron las actividades de estudio, las formas de aprendizaje y hábitos de estudio como tres clases para determinar un perfil académico de estudiantes a través de métodos de ensamble de clasificadores usando una red neuronal con tres clases: excelente, bueno y regular. Se obtuvo una precisión global de 80.4% para las tres clases.

En este trabajo se implementaron dos clasificadores de aprendizaje automático supervisado que fueron una red neuronal multicapa (MLP) y potenciación del gradiente (GB) para predecir el rendimiento escolar en español y matemáticas de los alumnos de la prueba ENLACE en los estados de Tlaxcala, Guerrero y Nuevo León y comparar los resultados obtenidos. Se consideraron variables como sexo, turno de la escuela, becario, el tipo de escuela y la marginación para determinar su influencia en la clasificación de los alumnos que estuvieron en la prueba durante cuatro años consecutivos, desde 6° de primaria en 2008 hasta 3° de secundaria en 2011. Se tomaron tres clases o categorías referenciadas a los niveles de logro, 0: insuficiente, 1: elemental, 2; bueno o excelente, en concordancia con la SEP.

## **2.4 MATERIALES Y MÉTODOS**

### **2.4.1 Colecta y preparación de datos**

De la información disponible de la prueba ENLACE se seleccionaron los datos de Tlaxcala, Guerrero y Nuevo León por tratarse de tres entidades del país pertenecientes a diferentes regiones y con sistemas educativos distintos. Se consideró a aquellos alumnos que estuvieron en la prueba 4 años consecutivos desde 6° de primaria hasta 3° de secundaria junto con variables que contextualizan el entorno del alumno y de la escuela tales como: turno, sostenimiento, modalidad o tipo de escuela, marginación y tamaño de la localidad, etcétera. A través de los formatos F911 (formato de registros administrativos) de la SEP y mediante la página del INEGI, se obtuvo información adicional que caracteriza al entorno escolar y la ubicación de la escuela.

En ambas entidades se seleccionó el subconjunto de alumnos que realizaron la prueba ENLACE durante 4 años consecutivos (2008 a 2011), de 6° de primaria hasta 3° de secundaria, que marcan el inicio de una trayectoria educativa (el paso de primaria hasta llegar a 3° de secundaria). Las variables que se consideraron como “predictoras” durante el desarrollo del análisis se muestran en el Cuadro 2.1.

**Cuadro 2.1. Variables contextuales y localización geográfica seleccionadas del Estado de Tlaxcala, Guerrero y Nuevo León, México.**

Variable	Descripción	Valores	Fuente de información
<i>n_esp</i>	Logro académico en español	0: insuficiente; elemental; 2: bueno o excelente	1: SEP, 2008
<i>n_mat</i>	Logro académico en matemáticas	0: insuficiente; elemental; 2: bueno o excelente	1:
<i>p_esp</i>	Puntaje en español	Rango entre 200 a 800 puntos	
<i>p_mat</i>	Puntaje en matemáticas	Rango entre 200 a 800 puntos	
<i>becario</i>	Condición de becario	0: no becario, 1: becario	
<i>turno</i>	Turno de la escuela	0: matutino, 1: vespertino	
<i>t_sost</i>	Tipo de sostenimiento	0: público, 1: privado	
<i>t_loc</i>	Tipo de la localidad	0: urbano, 1: rural	
<i>sexo</i>	Sexo	0: hombre, 1: mujer	
<i>t_esc</i>	Tipo de escuela	1: general, 2: indígena, 3: conafe, 4: particular, 5: telesecundaria, 6: técnica	
<i>t_loc</i>	Tamaño de la localidad	1: menos de 100 hab, 2: 100 hab a 249 hab, 3: 250 hab a 499 hab, 4: 500 hab a 2499 hab, 5: 2500 hab a 14 999 hab, 6: 15 000 y más hab.	
<i>n_mar</i>	Nivel de marginación	1: muy alto, 2: alto, 3: medio, 4: bajo, 5: muy bajo	
<i>lat</i>	Latitud	Decimal	INEGI, 2021

Variable	Descripción	Valores	Fuente de información
<i>lon</i>	Longitud	Decimal	
<i>alt</i>	Altitud	Metros sobre el nivel del mar	
<i>pob</i>	Población	En miles de habitantes	

**Fuente:** Elaboración propia con base en los formatos F911 de la SEP, información de ENLACE y con datos del Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Localidades del INEGI.

En 2008 se evaluaron 24 875 estudiantes del estado de Tlaxcala, de los cuales 11 036 estaban en 6° de primaria; Guerrero cuenta con información de 32 382 alumnos de 6° de primaria, que representa 41.5% de un total de 77 964 alumnos evaluados en la prueba ENLACE en 2008. Para Nuevo León, en 2008 se evaluaron 78 538 alumnos de 6° de primaria y de ese total se cuenta con 40 935 registros de aquellos estudiantes que estuvieron en 4 años de la prueba y que representan 52.1% del total de alumnos evaluados.

**Cuadro 2.2. Frecuencia de los niveles de logro de desempeño en español y matemáticas en Tlaxcala para una población de 11 036 estudiantes vigentes en 2008-2011.**

Año y asignatura evaluada	Niveles de logro (clases)		
	0 <sup>†</sup>	1 <sup>¶</sup>	2 <sup>§</sup>
ESP2008	1274	5949	3813
ESP2011	4194	5362	1480
MAT2008	1544	6039	3453
MAT2011	6254	3608	1174

<sup>†</sup>0: insuficiente; <sup>¶</sup>1: elemental; <sup>§</sup>2: bueno o excelente.

**Fuente:** Elaboración propia.

**Cuadro 2.3. Frecuencia de los niveles de logro de desempeño en español y matemáticas en Guerrero para una población de 32 382 estudiantes vigentes en 2008-2011.**

Año y asignatura evaluada	Niveles de logro (clases)		
	0 <sup>†</sup>	1 <sup>¶</sup>	2 <sup>§</sup>
EESP2008	7534	17 023	7825
MAT2008	6867	17 988	7527
ESP2011	14 401	13 261	4720
MAT2011	17 813	8210	6359

<sup>†</sup>0: insuficiente; <sup>¶</sup>1: elemental; <sup>§</sup>2: bueno o excelente.

**Fuente: Elaboración propia.**

**Cuadro 2.4. Frecuencia de los niveles de logro de desempeño en español y matemáticas en Nuevo León para una población de 40 935 estudiantes vigentes en 2008-2011.**

Año y asignatura evaluada	Niveles de logro (clases)		
	0 <sup>†</sup>	1 <sup>¶</sup>	2 <sup>§</sup>
ESP2008	6033	22 510	12 392
MAT2008	6952	23 750	10 233
ESP2011	17 604	18 185	5146
MAT2011	24 758	11 108	5069

<sup>†</sup>0: insuficiente; <sup>¶</sup>1: elemental; <sup>§</sup>2: bueno o excelente.

**Fuente: Elaboración propia.**

Las variables categóricas como tamaño de localidad, marginación y modalidad o tipo de escuela, se transformaron a variables indicadoras, resultado de esto se obtienen 22 variables predictoras. Para el análisis se generaron cuatro grupos de datos: español 2008-ESP2008, matemáticas 2008-MAT2008, español 2011-ESP2011 y matemáticas 2011-MAT2011 y en cada conjunto de datos las etiquetas de clase corresponden al nivel de logro de los alumnos obtenido en la materia correspondiente, *n\_esp* o *n\_mat*.

Para el procesamiento y formación de los cuatro conjuntos de datos se usó el software Statistical Analysis Software 9.4 (SAS). Para correr los algoritmos de aprendizaje automático se usó el software Python 3.8 a través de la biblioteca Scikit-learn con sus diferentes módulos para la ejecución de los códigos y la generación de resultados.

#### **2.4.2 Análisis exploratorio de los datos**

En México, todas las entidades han creado organismos institucionales encargados de la educación de su población, la gestión de sus leyes de educación, así como la contextualización regional del currículo. Fernández (2003), menciona la posibilidad de un efecto específico de la desigualdad de desarrollo económico de las entidades del país sobre los resultados académicos de los estudiantes, así mismo las desigualdades estructurales existentes en cada entidad inciden en los procesos de aprendizaje y el logro académico de los estudiantes. Las entidades federativas seleccionadas pertenecen a diferentes regiones del país que tienen condiciones económicas y estructurales diferenciadas. Bassols (1979) clasifica las regiones económicas de México y ubica a Tlaxcala en la región centro-este, Guerrero en la región este y a Nuevo León en la región noreste.

En 2008 se registraron 2 057 379 estudiantes de 6° de primaria que hicieron la prueba ENLACE. De éstos 24 875 fueron de Tlaxcala, 77 964 pertenecen a Guerrero y 78 538 son de Nuevo León. En 2011 hubo un total de 1 676 446 alumnos de 3° de secundaria que hicieron la prueba. En este año Tlaxcala registró 21 204 estudiantes, Guerrero a 57 512 alumnos y Nuevo León 69 161. De ambos años se seleccionaron a aquellos alumnos que estuvieron en la prueba de 2008 a 2011, es decir, los alumnos “sobrevivientes” o que lograron continuar sus estudios en una trayectoria educativa “ideal”, esto se hace para contrastar la información acerca de los niveles de logro en los que se ubicaron los estudiantes al concluir su educación primaria y secundaria como puntos de tránsito de nivel educativo.

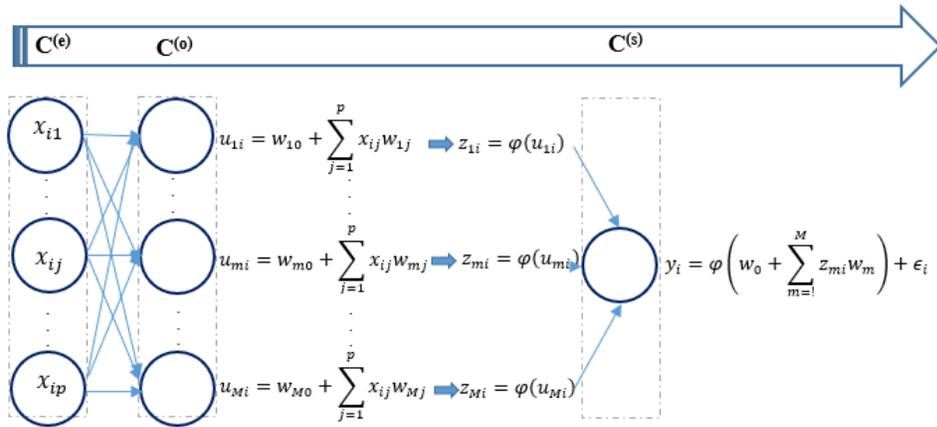
La teoría educativa señala que en general las mujeres suelen tener puntajes más elevados en español, mientras que los hombres sobresalen en matemáticas. Chen y Desjardins, (2008) mencionan que los apoyos a los estudiantes (becas escolares), se relacionan con la brecha de deserción escolar además de reducir las desigualdades.

#### **2.4.3 Clasificadores de aprendizaje automático**

Se implementaron dos algoritmos de aprendizaje automático supervisado para clasificación y un algoritmo para determinar la importancia de las variables “predictoras” o de entrada.

### 2.4.3.1 Red neuronal multicapa

La red neuronal multicapa (MLP, multilayer perceptron), es un aproximador universal y es muy eficiente para la aproximación de funciones de gran dimensión. MLP es además una red de retroalimentación con una o más capas de unidades entre la capa de entrada y la capa de salida. Las unidades de salida representan un hiperplano en el espacio de los patrones de salida.



$C^{(e)}$  = Capa de entrada;  $C^{(o)}$  = Capa oculta;  $C^{(s)}$  = Capa de salida

**Figura 2.1. Arquitectura de un clasificador de red neuronal multicapa (MLP). Fuente: Elaboración propia con base en González-Camacho et al., 2012.**

En la red neuronal todas las capas están conectadas. Primero se activa la capa oculta en donde se infieren  $M$  funciones  $\{z_i\}_{i=1; m=1}^{i=n; m=M}$ , con las variables de entrada  $\mathbf{x}_i = x_{i1}, \dots, x_{ip}$ , ( $j = 1, \dots, p$ ) se combinan usando la función lineal  $u_{mi} = w_{m0} + \sum_{j=1}^p x_{ij} w_{mj}$  con ( $m = 1, \dots, M$ ) y subsecuentemente se transforman usando una función no lineal de activación  $\varphi_m(\cdot)$  para obtener  $M$  puntajes  $z_{mi} = \varphi(u_{mi})$ . Los puntajes obtenidos se usan en la capa de salida como base para regresar una respuesta usando una función lineal de activación  $y_i = \varphi(w_0 + \sum_{m=1}^M z_{mi} w_m) + \epsilon_i$  (González-Camacho *et al.*, 2012).

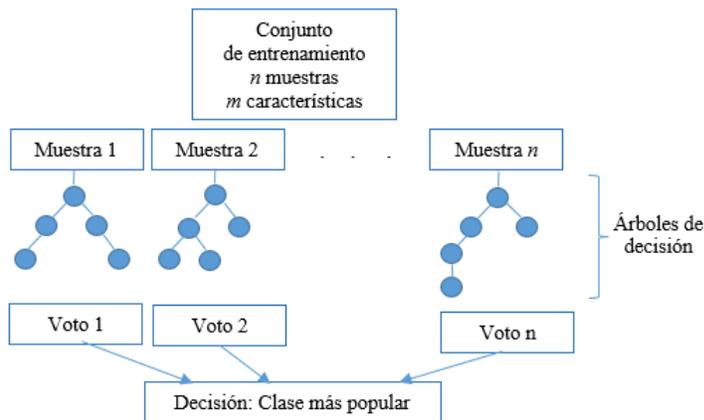
### 2.4.3.2 Clasificador potenciación del gradiente

El clasificador potenciación del gradiente (GB, gradient boosting) consiste en un conjunto de árboles que se van agregando secuencialmente en donde cada árbol va corrigiendo a su antecesor; cada árbol de decisión se entrena con los errores de su antecesor (Géron, 2019). Se pueden usar distintas funciones de pérdida como la suma de cuadrados del error, la función Laplaciana, etcétera. Una de las principales diferencias entre los métodos convencionales de

aprendizaje automático y los clasificadores de potenciación del gradiente es la optimización del espacio funcional, es decir, se parametriza la función  $\hat{f}$  en la forma funcional aditiva  $\hat{f}(x) = \hat{f}^M(x) = \sum_{i=0}^M \hat{f}_i(x)$ , donde  $M$  es el número de iteraciones,  $\hat{f}_0$  es la estimación inicial y  $\{\hat{f}_i\}_{i=1}^M$  son los incrementos de la función también llamados “boosts” o impulsos. En cada iteración se debe especificar la tasa de aprendizaje óptima (Natekin y Knoll, 2013).

### 2.4.3.3 Bosque aleatorio

Los bosques aleatorios (RF, random forest) son un conjunto de árboles de decisión entrenados generalmente a través del método de bagging, es decir, se combinan los resultados de varios clasificadores cada uno modelado con diferentes subconjuntos tomados de la misma población para reducir la varianza de las predicciones. El algoritmo RF promedia las predicciones sobre muchos árboles individuales. Los árboles individuales se basan en muestras con reemplazo o bootstrap (Schonlau y Yuyan, 2020). Cada árbol crece y las nuevas instancias se predicen a partir de agregaciones de los árboles (por mayoría de votos).



**Figura 2.2. Construcción de un bosque aleatorio. Fuente: Elaboración propia.**

Otra aplicación de los bosques aleatorios es que facilitan la medición de la importancia relativa de las características del conjunto de datos. A través de la biblioteca de Scikit-learn de Python, se mide la importancia de una característica mediante la reducción de la impureza en los nodos de los árboles que usan esa característica. Es decir, es un promedio ponderado donde el peso de cada nodo es igual al número de muestras de entrenamiento asociadas a él (Géron, 2019).

#### 2.4.4 Criterios de desempeño de los modelos de predicción

Para evaluar los modelos de clasificación MLP, GB y RF las métricas se obtienen a partir de una matriz de confusión (*MC*) (Figura 2.3). Las filas representan número de muestras en la clase observada y las columnas el número de predicciones de cada clase. La diagonal principal de *MC* corresponde al número de muestras que el algoritmo clasifica correctamente en cada clase. Si la *MC* solo tiene valores positivos en la diagonal, indica que el clasificador clasifica correctamente todas las muestras. En la *MC* se describen el conteo de los verdaderos positivos (*VP*), verdaderos negativos (*VN*), falsos positivos (*FP*) y falsos negativos (*FN*).

Matriz de confusión		Clases predichas		
		Positivo (P)	Negativo (N)	
Clases reales	Positivo (P)	<i>VP</i>	<i>FN</i>	<b>Precisión:</b> Porcentaje de predicciones positivas correctas $P=VP/VP+FP$
	Negativo (N)	<i>FP</i>	<i>VN</i>	
		Sensibilidad (recall): Porcentaje de casos positivos	Especificidad (especificity): porcentaje de casos negativos	Accuracy o precisión global: porcentaje de predicciones correctas
		$S=VP/VP+FN$	$VN/VN+FP$	$PG=VP+VN/VP+VN+FN+FP$

**Figura 2.3. Matriz de confusión y métricas de desempeño. Fuente: elaboración propia.**

Otra métrica importante es el valor de  $F1-Score = 2 \frac{P \times S}{P+S}$  que resume la *P* y la *S* en una sola métrica. Usualmente se usa cuando la distribución de las clases es desigual. Con esta métrica se pueden obtener cuatro casos posibles para cada clase 1) alta *P* y alta *S*, 2) alta *P* y baja *S*, 3) baja *P* y alta *S*, 4) baja *P* y baja *S*. La curva ROC (receiver operating characteristics) grafica la *S* versus  $1-E$  y se construye con base en la unión de diferentes puntos de corte en donde el eje Y es la *S* y el eje X es la  $1-E$ . El área por debajo de la curva es el  $AUC_{ROC}$  (área bajo la curva), que indica que tan buena es la clasificación, un valor cercano a 0.5% indica

una mala clasificación (Cerde y Cifuentes, 2011). La curva *P-S* es el resultado de graficar la *P* en el eje Y y *S* en el eje X. El punto más bajo (0,0) no logra hacer distinciones en las clases y el punto más alto (1,1) significa que tanto la *P* y *S* son altas por lo que el modelo hace una clasificación perfecta. El objetivo es que la curva se acerque lo más posible a (1,1) lo que es una buena *P* y *S*. La curva genera un área bajo la curva  $AUC_{P-S}$  que entre más cercano a 1 sea, mejor será el modelo (Saito y Rehmsmeier, 2015).

#### 2.4.5 Entrenamiento y validación de los modelos

El entrenamiento de MLP y GB se realizó en dos etapas, la primera es la selección de los hiperparámetros óptimos y la segunda es la evaluación de ambos modelos para obtener las métricas de desempeño. Cada conjunto de datos se dividió en dos, 80.0% para muestras de entrenamiento y 20.0% prueba. Del conjunto de datos se fusionaron los niveles 2 y 3 para equilibrar las clases y mejorar la clasificación de los modelos. A los conjuntos de datos, (ESP2008, MAT2008, ESP2011 y MAT2011) se aplicó el clasificador RF para seleccionar aquellas características que fueran más importantes.

**Cuadro 2.5. Muestras de prueba por clase, Tlaxcala.**

Clase	ESP2008	MAT2008	ESP2011	MAT2011
0 <sup>†</sup>	255	309	839	1251
1 <sup>¶</sup>	1190	1208	1073	722
2 <sup>§</sup>	763	691	296	235

<sup>†</sup>0: insuficiente; <sup>¶</sup>1: elemental; <sup>§</sup>2: bueno o excelente.

**Fuente: Elaboración propia con base en los datos de ENLACE 2008-2011.**

**Cuadro 2.6. Muestras de prueba por clase, Guerrero.**

Clase	ESP2008	MAT2008	ESP2011	MAT2011
0 <sup>†</sup>	1507	1373	2881	3563
1 <sup>¶</sup>	3405	3598	2652	1642
2 <sup>§</sup>	1565	1506	944	1272

<sup>†</sup>0: insuficiente; <sup>¶</sup>1: elemental; <sup>§</sup>2: bueno o excelente.

**Fuente: Elaboración propia con base en los datos de ENLACE 2008-2011.**

**Cuadro 2.7. Muestras de prueba por clase, Nuevo León.**

Clase	ESP2008	MAT2008	ESP2011	MAT2011
0	1207	1390	3521	4951
1	4502	4750	3637	2222
2	2478	2047	1029	1014

†0: insuficiente; †1: elemental; †2: bueno o excelente.

Fuente: Elaboración propia con base en los datos de ENLACE 2008-2011.

#### 2.4.5.1 Selección de hiperparámetros óptimos

El ajuste de hiperparámetros se hizo mediante búsqueda en malla explorando diferentes valores que se definieron para cada hiperparámetro (Cuadro 2.8). Se usa 80.0% de los datos para el ajuste de hiperparámetros utilizando CV con cinco particiones (subconjuntos). En CV se hacen  $k$  particiones sin reemplazo del conjunto de datos de entrenamiento. El modelo se entrena usando  $k-1$  subconjuntos para entrenamiento y la restante para prueba. El procedimiento se repite hasta que cada uno de los  $k$  subconjuntos haya sido usado como conjunto de validación, se obtienen  $k$  modelos con estimaciones de desempeño (Raschka, 2018).

**Cuadro 2.8. Intervalo de valores para la prueba de hiperparámetros para Tlaxcala, Guerrero y Nuevo León.**

Modelo	Hiperparámetros	Intervalo de valores
MLP	<i>nco</i>	50, 100, 200, 250
	<i>fa</i>	Tanh, relu, logistic
	<i>op</i>	Sgd, Adam, lbfgs
	<i>re</i>	0.0001, 0.001, 0.1
	<i>ta</i>	Constant, adaptative
GB	<i>mi</i>	90, 100, 110
	<i>pa</i>	4, 5, 6
	<i>ha</i>	3, 4

*nco*: neuronas en capa oculta; *fa*: función de activación; *op*: optimizador de pesos; *re*: regularizador; *ta*: tasa de aprendizaje; *mi*: máximo de iteraciones; *pa*: profundidad de árbol; *ha*: hojas por árbol. Fuente: Elaboración propia.

#### **2.4.5.2 Evaluación del desempeño de los modelos de clasificación**

Después de que se obtienen los valores óptimos de los hiperparámetros, el modelo se entrenó considerando los datos completos de cada conjunto de datos y se estimó la *PG*. A través de validación cruzada con  $k = 5$  particiones y usando los valores óptimos de hiperparámetros, se estima una precisión y para obtener la *PG* del modelo, al finalizar las iteraciones, se hace un promedio para tener un único valor de *PG*.

#### **2.4.6 Importancia relativa de las características de entrada**

La importancia de variables es utilizada en los modelos de aprendizaje automático y trata de ver cómo se afecta la salida del modelo cuando se hacen cambios en las variables de entrada. Aquellas variables predictoras que más variabilidad producen en la salida son las que más influencia tienen y por lo tanto las que mejor explican el modelo.

Para determinar la importancia de las variables predictoras se usa RF. En este modelo se usa bootstrap y en dicho proceso se deja aproximadamente un tercio de los casos de la muestra; los casos no considerados para entrenar el árbol se llaman out of-bag (OOB) y con éstos se puede estimar un error de clasificación. Primero se escoge un OOB, después se toma una variable al azar y se permutan sus valores dentro de los datos de entrenamiento para que esta variable descorrelacione lo que el modelo aprendió. Posteriormente se vuelve a calcular el OOB para compararlo con el error inicial. Si el error cambia entonces dicha variable es importante. El proceso se repite con todas las variables y al final se ordenan de acuerdo con los cambios que produjeron cada una en los errores OOB (Medina y Ñique, 2017).

La determinación de la importancia relativa de las 22 variables predictoras se hizo con RF. Mediante validación cruzada con cinco particiones se determinaron los valores óptimos de los hiperparámetros del modelo. Posteriormente se ajusta el modelo para determinar la importancia de las variables a través de la biblioteca Scikit-learn de Python con feature importance. Este procedimiento se realizó para los cuatro conjuntos de datos: ESP2008, MAT2008, ESP2011 y MAT2011.

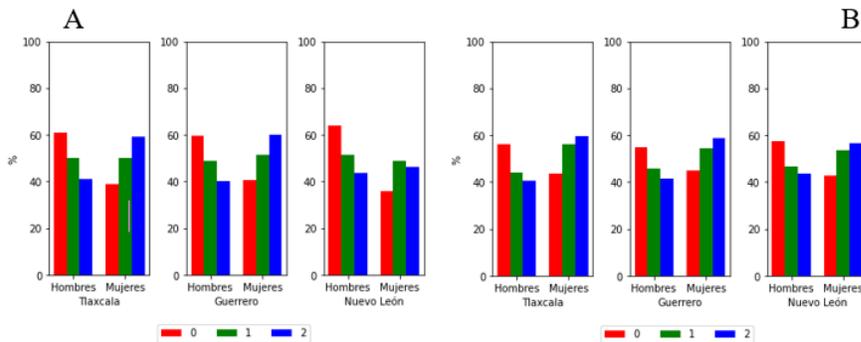
## 2.5 RESULTADOS Y DISCUSIÓN

Para la generación de los resultados presentados en esta sección, se elaboraron scripts en Python que se describen en el Anexo.

### 2.5.1 Análisis exploratorio

- **Español**

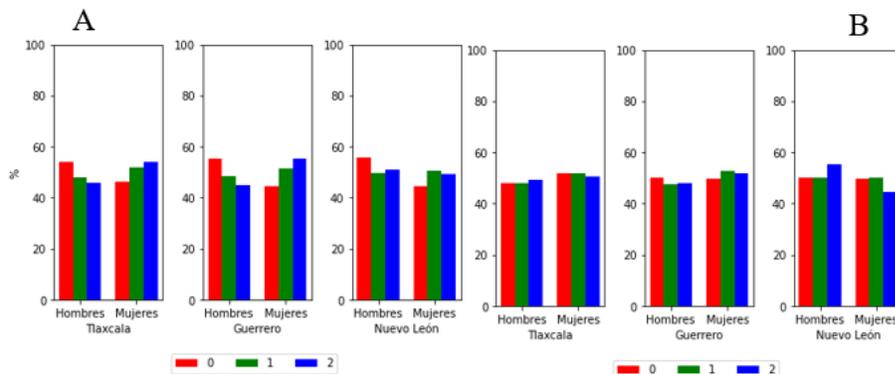
De la Figura 2.4, en español 2008, en Guerrero, las mujeres son las que se encuentran en mayor proporción en los niveles más altos, 51.3% en el nivel de logro 1-elemental y 60.0% en el nivel 2-bueno o excelente. En Nuevo León en esta misma materia, este comportamiento solo se observa en el nivel 2, con 56.3% de mujeres contra 43.6% de hombres.



**Figura 2.4.** Nivel de logro de los estudiantes de Tlaxcala, Guerrero y Nuevo León por sexo. A) Español 2008, B) Español 2011. Fuente: Elaboración propia.

- **Matemáticas**

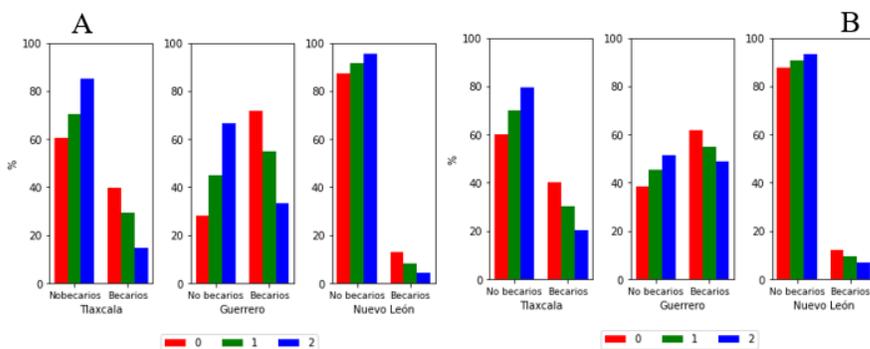
En matemáticas, para Guerrero y Nuevo León, se observa que más mujeres se encuentran en los niveles superiores de aprovechamiento, 1 y 2. Pero en Nuevo León son más hombres.



**Figura 2.5.** Nivel de logro de los estudiantes de Tlaxcala, Guerrero y Nuevo León por sexo. A) Matemáticas 2008, B) Matemáticas 2011. Fuente: Elaboración propia.

Del grupo de variables predictoras también se encuentra la condición de becario (*becario*). Esta variable proviene de las becas y apoyos económicos que los estudiantes tuvieron a lo largo de su paso por la escuela. De la figura 2.6 se observa que de los resultados de español 2008 en las tres entidades, los no becarios son los que obtienen mejores resultados en la prueba. En 2011 en español, sin embargo, para Guerrero se observa que los becarios tuvieron mayor proporción en los niveles 1 y 2 con 54.9% y 53.2% respectivamente.

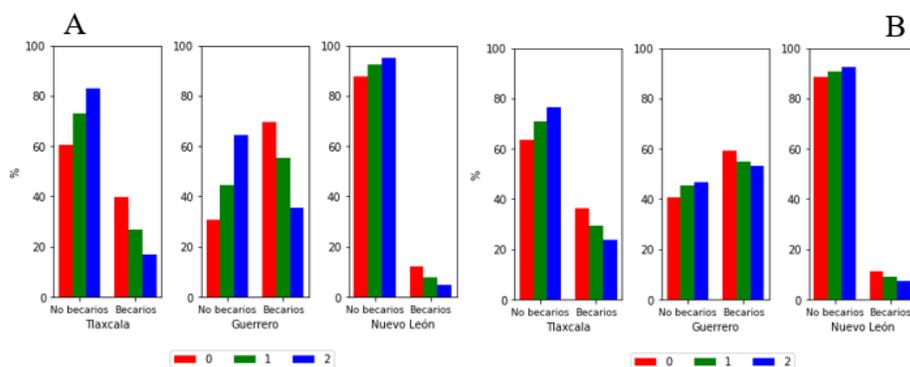
- **Español**



**Figura 2.6. Nivel de logro de los estudiantes de Tlaxcala, Guerrero y Nuevo León por condición de becario A) Español 2008, B) Español 2011. Fuente: Elaboración propia.**

- **Matemáticas**

En Guerrero, los becarios son más representativos en el nivel 1 con 54.7%, sin embargo, este grupo es el que tiene más alumnos en el nivel 0 (61.7%). En Nuevo León, en los dos años, los alumnos que no tuvieron beca son los que se encuentran proporcionalmente en niveles superiores de logro (1 y 2), de igual proporción tienen alumnos en el nivel 0 de logro.



**Figura 2.7. Nivel de logro de los estudiantes de Tlaxcala, Guerrero y Nuevo León por condición de becario. A) Matemáticas 2008, B) Matemáticas 2011. Fuente: Elaboración propia.**

## 2.5.2 Hiperparámetros óptimos

En el Cuadro 2.9 se presentan los hiperparámetros que se ajustaron a los dos modelos y los valores óptimos que se obtuvieron mediante validación cruzada  $CV = 5$ . y búsqueda en retícula.

**Cuadro 2.9. Hiperparámetros óptimos para el clasificador red neuronal multicapa (MLP) y potenciación del gradiente (GB) y cada conjunto de datos (asignatura y año), Tlaxcala.**

Modelo	Hiperparámetros	ESP2008	ESP2011	MAT2008	MAT2011
MLP	<i>nco</i>	100	200	50	50
	<i>fa</i>	Logistic	Tanh	Relu	Logistic
	<i>op</i>	Adam	Sgd	Sgd	Adam
	<i>re</i>	0.1	0.1	0.1	0.0001
	<i>ta</i>	Constant	Adaptative	Constant	Constant
GB	<i>mi</i>	90	90	90	110
	<i>pa</i>	4	4	4	4
	<i>ha</i>	3	4	3	4

*nco*: neuronas en capa oculta; *fa*: función de activación; *op*: optimizador de pesos; *re*: regularizador; *ta*: tasa de aprendizaje; *mi*: máximo de iteraciones; *pa*: profundidad de árbol; *ha*: hojas por árbol. **Fuente: Elaboración propia.**

**Cuadro 2.10. Hiperparámetros óptimos para el clasificador red neuronal multicapa (MLP) y potenciación del gradiente (GB) y cada conjunto de datos (asignatura y año), Guerrero.**

Modelo	Hiperparámetros	ESP2008	ESP2011	MAT2008	MAT2011
MLP	<i>nco</i>	100	50	250	200
	<i>fa</i>	Relu	Relu	Relu	Logistic
	<i>op</i>	Sgd	Sgd	Adam	Adam
	<i>re</i>	0.1	0.1	0.1	0.0001
	<i>ta</i>	Adaptative	Adaptative	Constant	Constant
GB	<i>mi</i>	110	110	90	110
	<i>pa</i>	4	6	5	5
	<i>ha</i>	4	4	4	4

*nco*: neuronas en capa oculta; *fa*: función de activación; *op*: optimizador de pesos; *re*: regularizador; *ta*: tasa de aprendizaje; *mi*: máximo de iteraciones; *pa*: profundidad de árbol; *ha*: hojas por árbol. **Fuente: Elaboración propia.**

**Cuadro 2.11. Hiperparámetros óptimos para el clasificador red neuronal multicapa (MLP) y potenciación del gradiente (GB) y cada conjunto de datos (asignatura y año), Nuevo León.**

Modelo	Hiperparámetros	ESP2008	ESP2011	MAT2008	MAT2011
MLP	<i>nco</i>	300	50	250	200
	<i>fa</i>	Logistic	Relu	Relu	Logistic
	<i>op</i>	Adam	Sgd	Adam	Adam
	<i>re</i>	0.001	0.1	0.1	0.0001
	<i>ta</i>	Constant	Adaptative	Constant	Constant
GB	<i>mi</i>	100	110	100	110
	<i>pa</i>	6	4	5	6
	<i>ha</i>	3	4	3	4

*nco*: neuronas en capa oculta; *fa*: función de activación; *op*: optimizador de pesos; *re*: regularizador; *ta*: tasa de aprendizaje; *mi*: máximo de iteraciones; *pa*: profundidad de árbol; *ha*: hojas por árbol. **Fuente: Elaboración propia.**

### 2.5.3 Desempeño de los clasificadores

Sobre los resultados de los modelos, se presentan cuadros comparativos por materia evaluada y año de la prueba ENLACE. Para cada modelo se presenta la *PG* del modelo con su desviación estándar, la curva ROC y las curvas de *P-S*.

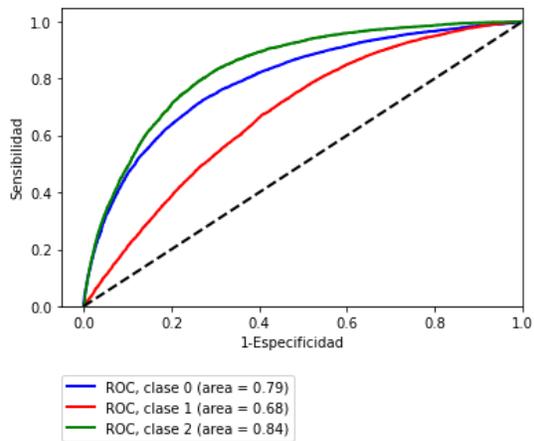
**Cuadro 2.12. Precisión global-*PG* de clasificación de los modelos por año y materia evaluada.**

Modelo	ESP2008	ESP2011	MAT2008	MAT2011
Tlaxcala				
MLP	0.701 (+- 0.033)	0.611 (+- 0.019)	0.672 (+- 0.034)	0.631 (+- 0.008)
GB	0.695 (+- 0.024)	0.610 (+- 0.016)	0.688 (+- 0.008)	0.635 (+- 0.006)
Guerrero				
MLP	0.671 (+- 0.011)	0.597 (+- 0.014)	0.677 (+- 0.007)	0.623 (+- 0.005)
GB	0.667 (+- 0.005)	0.607 (+- 0.011)	0.674 (+- 0.009)	0.634 (+- 0.003)
Nuevo León				
MLP	0.672 (+- 0.014)	0.612 (+- 0.012)	0.651 (+- 0.034)	0.657 (+- 0.005)
GB	0.690 (+- 0.008)	0.614 (+- 0.008)	0.663 (+- 0.029)	0.663 (+- 0.005)

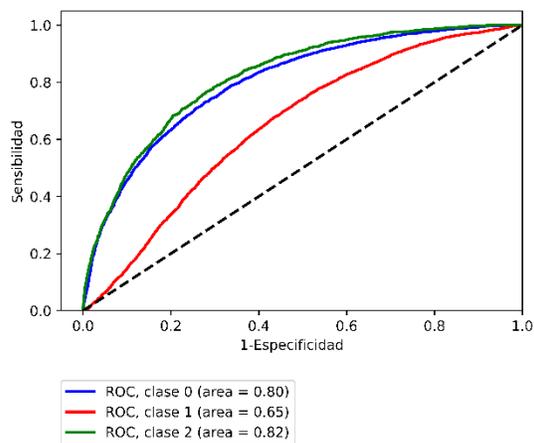
**Fuente: Elaboración propia.**

Al aplicar los dos algoritmos propuestos a las entidades seleccionadas para español, con MLP en Tlaxcala se obtuvo el mejor desempeño de clasificación para los alumnos de 6° de primaria en 2008, con *PG* de 70.1%. En matemáticas para 3° de secundaria en 2011, GB fue superior con *PG* de 63.5% de muestras clasificadas correctamente. En Guerrero, con MLP se alcanzó una *PG* de 67.1% y de las *P* por clases, la clase 1 tiene una *P* de 67.0% y la clase 2 de 76.0%. De la *MC*, casi 50.0% de las muestras de la clase 0 se clasifican como clase 1. En 2011 GB obtuvo los mejores resultados al presentar *PG* de 60.7%. Con la métrica área bajo la curva  $AUC_{ROC}$  para las clases 0 y 2 se tuvieron valores de 0.79 y 0.84 lo que se interpreta como una buena clasificación (Figura 2.8). Del Cuadro 2.10, MLP obtuvo mejor desempeño para clasificar la clase 2 con altos valores de *P*, *FI* y  $AUC_{P-S}$ , es decir, el modelo detecta muy bien la clase. En Nuevo León, en español, el modelo que obtuvo mejores valores fue GB con la métrica de *PG* = 69.0% de las muestras se clasificaron correctamente, en 2011 *PG* = 61.4%. De la Figura 2.9 se observa que las clases 0 y 2 son las que tiene una mayor  $AUC_{ROC}$  lo que es una buena clasificación. Con la métrica de área bajo la curva precisión-

sensibilidad ( $AUC_{P,S}$ ), estas clases tienen valores de 0.50 y 0.72 en 2008 y de 0.74 y 0.44 en 2011.



**Figura 2.8.** Curva ROC para el modelo GB en español 2011, Guerrero. Fuente: elaboración propia.



**Figura 2.9.** Curva ROC para el modelo GB en español 2011, Nuevo León. Fuente: elaboración propia.

**Cuadro 2.13. Valores de P y F1, AUC<sub>ROC</sub> y AUC<sub>P-S</sub> para cada modelo por año y clase en español, Guerrero.**

Modelo	Clase	ESP2008			ESP2011		
		<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>	<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>
MLP	0	0.56	0.85	0.60	0.72	0.78	0.70
	1	0.73	0.72	0.68	0.59	0.66	0.60
	2	0.70	0.91	0.78	0.50	0.83	0.44
GB	0	0.56	0.85	0.58	0.71	0.79	0.75
	1	0.73	0.71	0.68	0.60	0.68	0.55
	2	0.69	0.91	0.78	0.41	0.84	0.45

*P*: precisión; *FI*: f1-score; AUC<sub>ROC</sub>: Área bajo la curva ROC; AUC<sub>P-S</sub>: Área bajo la curva *P-S*. **Fuente:** Elaboración propia.

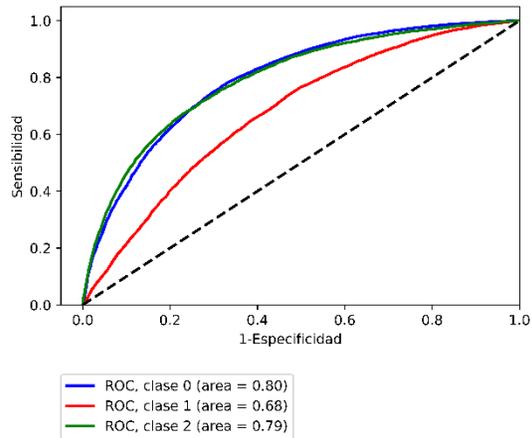
**Cuadro 2.14. Valores de P y F1, AUC<sub>ROC</sub> y AUC<sub>P-S</sub> para cada modelo por año y clase en español, Nuevo León.**

Modelo	Clase	ESP2008			ESP2011		
		<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>	<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>
MLP	0	0.41	0.85	0.50	0.69	0.80	0.74
	1	0.76	0.71	0.70	0.63	0.66	0.56
	2	0.71	0.86	0.71	0.48	0.84	0.45
GB	0	0.41	0.87	0.50	0.68	0.80	0.74
	1	0.76	0.71	0.70	0.62	0.65	0.56
	2	0.71	0.86	0.72	0.48	0.82	0.44

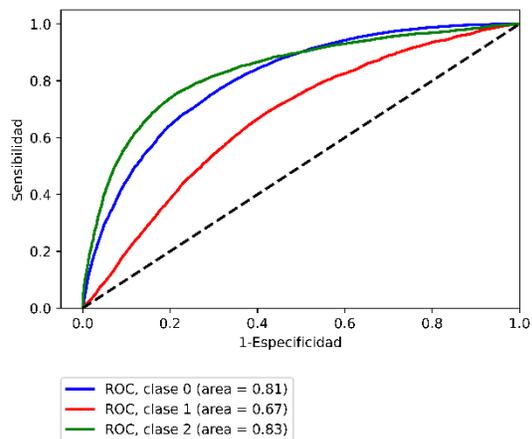
*P*: precisión; *FI*: f1-score; AUC<sub>ROC</sub>: Área bajo la curva ROC; AUC<sub>P-S</sub>: Área bajo la curva *P-S*. **Fuente:** Elaboración propia.

Para los resultados de clasificación en matemáticas, en Tlaxcala se obtuvo la mayor *PG* de las tres entidades, con 68.8% de muestras clasificadas correctamente con el modelo GB. En Guerrero, para matemáticas en 2008 el modelo que mejor clasificó las muestras fue MLP con *PG* de 67.7%, mientras que en 2011 el mejor modelo fue GB con *PG* de 63.4%. Para los dos modelos evaluados, las mayores *P* por clase en 2008 corresponden a la clase 1 y la 2, en tanto que en 2011 las mayores *P* son de las clases 0 y 2; esto está relacionado con el número de muestras y el cambio observado en ambos años. Para Nuevo León, en 2011 con GB se obtuvo una *PG* en matemáticas de 66.3%. Y de las precisiones por clase se tiene que las

clases 0 y 2 con las que tiene mayor  $P$ , esto en concordancia con los resultados que se obtuvieron del área bajo la curva  $AUC_{ROC}$  de la Figura 2.10 para dichas clases. Para esta entidad, de los valores observados del Cuadro 2.16, para matemáticas 2011, se observa que el modelo GB detecta bien la clase 0 al tener altos valores de  $FI$  y  $AUC_{P-S}$  (0.81).



**Figura 2.10.** Curva ROC para el modelo GB en matemáticas 2011, Guerrero. Fuente: elaboración propia.



**Figura 2.11.** Curva ROC para el modelo GB en matemáticas 2011, Nuevo León. Fuente: elaboración propia.

**Cuadro 2.15. Valores de P y F1, AUC<sub>ROC</sub> y AUC<sub>P-S</sub> para cada modelo por año y clase en matemáticas, Guerrero.**

Modelo	Clase	MAT2008			MAT2011		
		<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>	<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>
MLP	0	0.49	0.84	0.56	0.78	0.79	0.81
	1	0.74	0.71	0.72	0.39	0.66	0.37
	2	0.68	0.90	0.77	0.57	0.82	0.54
GB	0	0.50	0.84	0.55	0.78	0.80	0.82
	1	0.75	0.71	0.71	0.33	0.68	0.39
	2	0.68	0.90	0.77	0.54	0.79	0.55

*P*: precisión; *FI*: f1-score; AUC<sub>ROC</sub>: Área bajo la curva ROC; AUC<sub>P-S</sub>: Área bajo la curva *P-S*. **Fuente:** Elaboración propia

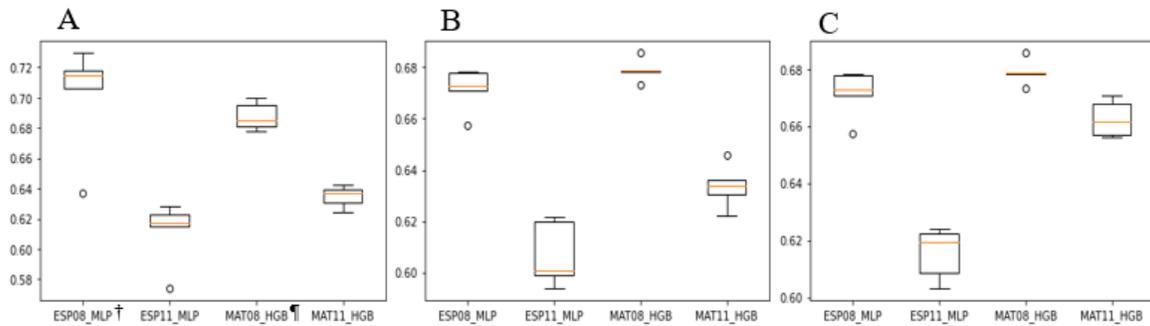
**Cuadro 2.16. Valores de P y F1, AUC<sub>ROC</sub> y AUC<sub>P-S</sub> para cada modelo por año y clase en matemáticas, Nuevo León.**

Modelo	Clase	MAT2008			MAT2011		
		<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>	<i>FI</i>	AUC <sub>ROC</sub>	AUC <sub>P-S</sub>
MLP	0	0.46	0.85	0.51	0.82	0.80	0.85
	1	0.77	0.67	0.69	0.37	0.66	0.38
	2	0.67	0.84	0.61	0.53	0.84	0.46
GB	0	0.41	0.85	0.50	0.81	0.81	0.85
	1	0.77	0.68	0.71	0.39	0.67	0.38
	2	0.67	0.86	0.66	0.51	0.83	0.47

*P*: precisión; *FI*: f1-score; AUC<sub>ROC</sub>: Área bajo la curva ROC; AUC<sub>P-S</sub>: Área bajo la curva *P-S*. **Fuente:** Elaboración propia

#### 2.5.4 Comparación de los modelos de clasificación

Como se planteó en el inicio del trabajo, se aplicaron dos modelos de aprendizaje automático supervisado a los cuatro grupos de datos generados para las tres entidades del país. En la Figura 2.12 se presentan los resultados obtenidos aplicando MLP y GB. Se observa que, en las tres entidades, en español 2008 se obtienen mejores resultados con MLP. En matemáticas, se obtienen las mejores precisiones con GB.

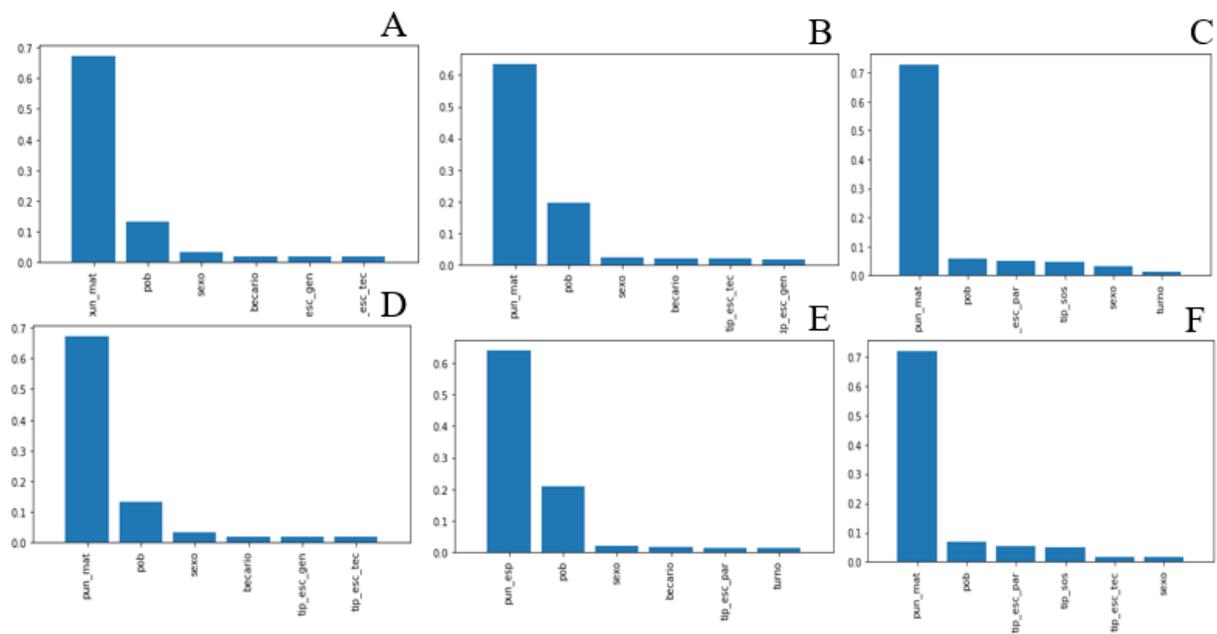


†MLP: Red neuronal multicapa; †HGB: Potenciación del gradiente.

**Figura 2.12. Comparación de precisión global para los modelos por entidad federativa. A) Tlaxcala, B) Guerrero, C) Nuevo León. Fuente: Elaboración propia.**

### 2.5.5 Importancia relativa de características

Al aplicar RF a los cuatro grupos de datos, para las dos entidades, se encontró que en español y matemáticas 2008 y 2011 la variable de calificación,  $p\_esp$ ,  $p\_mat$ , son las más importantes (Figura 2.13). En esta parte también se observa la importancia de otras variables como la  $pob$ ,  $sexo$ ,  $becario$  y  $t\_esc$ . sin embargo, evaluando los dos modelos seleccionados y considerando estas variables y siete más ( $pob$ ,  $lat$ ,  $lon$ ,  $alt$ ,  $sexo$ ,  $becario$ ,  $turno$ ) la PG de los modelos de clasificación no mejoró, y se decidió considerar todo el conjunto de variables y la partición propuesta de 80.0% y 20.0%.



**Figura 2.13. Importancia de seis variables con RF. Español 2011 A) Tlaxcala, B) Guerrero C) Nuevo León. Matemáticas 2011 D) Tlaxcala E), Guerrero F) Nuevo León. Fuente Elaboración propia**

### 2.5.6 Discusión

A través del desarrollo del trabajo se observó que las variables más importantes en la clasificación de los modelos fueron *p\_esp* y *p\_mat*. En la prueba ENLACE estas variables se usan para hacer los cortes que identifican a que clase pertenecen los alumnos. Aunque las tres entidades tienen sistemas educativos diferentes, las variables de localización geográfica de las escuelas y de *población* son las que más influyen en la predicción de las clases, aunque posteriormente se incluyan todas las variables a los modelos. El *sexo* como se mencionó anteriormente influye en los resultados que los alumnos obtienen, aunque con los resultados obtenidos, no es determinante en la clasificación. La variable *becario* también se consideró como importante pero no se observan diferencias importantes en los resultados por entidad federativa. En las tres entidades se obtuvieron mejores clasificaciones en 2008; con MLP para español y GB en matemáticas.

## 2.6 CONCLUSIONES

Con base en los objetivos planteados y resultados obtenidos, de los dos algoritmos de clasificación aplicados a las muestras de alumnos de 6° de primaria y 3° de secundaria en Guerrero y Nuevo León se obtuvo que el GB es el que mejor clasifica a los alumnos en alguna de las tres clases correspondientes a los niveles de logro. En Nuevo León GB fue el algoritmo que mejores resultados tuvo, con las muestras de 2008, en español se tuvo una *PG* de 69.0% y de 66.3% en matemáticas. En 2011 este mismo algoritmo alcanzó una *PG* de 61.4% en español y de 66.3% en matemáticas. Para Guerrero en 2008 MLP fue el mejor clasificador, 67.1% de *PG* en español y 67.7% en matemáticas; en 2011 con GB se obtuvo una *PG* mayor, 60.7% en español y 63.4% en matemáticas. Acerca de las variables contextuales, en ambas entidades los puntajes en español y matemáticas tuvieron la mayor importancia ( $p_{esp}$ ,  $p_{mat}$ ), seguidas de la variable población; también resultaron importantes, aunque en menor medida fueron el sexo, y en Guerrero el tipo de sostenimiento de las escuelas a las que acuden los estudiantes, ya sean escuelas públicas o privadas.

Resulta importante considerar la selección de diferentes variables que pueden influir en la clasificación de los alumnos en un nivel de logro académico. En México la deserción escolar es uno de los principales problemas con los que se enfrentan las escuelas, de esta forma si se logra identificar a tiempo a aquellos estudiantes con bajo desempeño es posible que se puedan sugerir políticas o programas para su atención oportuna. Este trabajo se puede complementar considerando variables diferentes o adicionales a las que ya se tomaron y ver el desempeño de los modelos de clasificación propuestos.

## CONCLUSIONES GENERALES

De acuerdo con los objetivos planteados en la investigación y de los resultados obtenidos, se concluye que:

Con la implementación de los modelos de aprendizaje automático supervisado, bosque aleatorio, red neuronal multicapa y potenciación del gradiente, para clasificar a los alumnos que hicieron la prueba ENLACE en cuatro años consecutivos, en tres clases o niveles de logro 0: insuficiente, 1: elemental, 2: bueno o excelente (español y matemáticas 2008 y 2011). Se determinó que conocer los puntajes o calificaciones que un estudiante tiene en español tienen la mayor importancia relativa para predecir la clasificación en matemáticas y viceversa. Variables contextuales como población, sexo, tipo de escuela (general, indígena, comunitaria, etcétera) o si la escuela es pública o privada, también influyen en la clasificación de los alumnos en alguno de los niveles de logro.

Al aplicar los algoritmos de aprendizaje automático a las entidades seleccionadas, Tlaxcala, Guerrero y Nuevo León, se obtuvo que, para español, con la red neuronal multicapa se obtiene mejor precisión global de clasificación. En matemáticas, potenciación del gradiente fue el que mejor clasificó. En 2008 los algoritmos tuvieron mejor desempeño para clasificar las clases 1 y 2 y en 2011 lo fueron para las clases 0 y 2. Para la comparación de los niveles de logro entre entidades, se observó que las mujeres se encuentran en mayor proporción en los niveles de logro 1 y 2, en ambas materias evaluadas. Al observar por la condición de becario, esta variable influye en los resultados de español y matemáticas en Guerrero ya que mayor proporción de becarios se ubican en el nivel de logro 1. En las demás entidades, los no becarios son los que se encuentran en mayor proporción en los niveles de logro 1 y 2. Los resultados obtenidos son de gran utilidad, sobre todo en cuestiones de investigación educativa. En diversos trabajos se ha hecho asociación entre variables externas a la escuela como influencia en el desempeño escolar tal como modelos mixtos o jerárquicos; aún son incipientes los que utilizan algoritmos de aprendizaje automático y datos a gran escala. Lo desarrollado en esta investigación da pauta a utilizar una metodología alterna que ayude a identificar qué variables son importantes en el grado de logro de los estudiantes, poder hacer una focalización de los que se identificaron con bajo desempeño y plantear una atención dirigida.

## LITERATURA CITADA

- Altabrawee, H., J. Osama A, and Q. Ajmi S. 2019. Predicting students' performance using machine learning techniques. *Journal of university of Babylon, pure and applied sciences*. 27: 194-205. Doi: <https://doi.org/10.29196/jubpas.v27i1.2108>
- Álvarez, J., V. García., y H. Patrions. 2007. Institutional Effects as Determinants of Learning Outcomes. Exploring State Variations in Mexico. Policy Research Working Paper 4286. Banco Mundial.
- Backhoff E. E., y S. Contreras R. 2014. Corrupción de la medida e inflación de los resultados de enlace. *Revista Mexicana de Investigación Educativa*. 19: 1267-1283.
- Bassols, A. 1979. Geografía, Subdesarrollo y Regionalización. Editorial Nuestro Tiempo. México Distrito Federal. 5ta edición. Avenida Copilco 300, México D. F. 253 p.
- Borkar, S., and K. Rajeswari. 2014. Attributes selection for predicting students' academic performance using education data mining and artificial neural network. *International journal of computer applications*. 86: 25-29. Doi: <https://doi.org/10.5120/15022-3310>
- Breiman, L. 2001. Random forests. *Machine learning*. 45: 5-32. Doi: <https://doi.org/10.1023/A:1010933404324>
- Carvallo P. M., J. Caso N., y L. A. Contreras N. 2007. Estimación del efecto de variables contextuales en el logro académico de estudiantes de Baja California. *Revista Electrónica de Investigación Educativa*. 9: 1-14.
- Cerda, J y L. Sifuentes. 2011. Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos. *Revista Chilena de Infectología*. 29: 138-141. Doi: <http://dx.doi.org/10.4067/S0716-10182012000200003>
- Chen, R. and S. L. Desjardins. 2008. Exploring the effects of financial aid on the gap in student dropout risks by income level. *Research in higher education*. 49: 1-18. Doi: <https://doi.org/10.1007/s11162-007-9060-9>
- Đambić, G., M. Krajcar., and D. Bele. 2016. Machine learning model for early detection of higher education students that need additional attention in introductory programming courses. *International journal of digital technology & economy*. 1: 11.
- De Hoyos, R. E., J. M. Espino., and V. García. 2012. Determinantes del logro escolar en México. Primeros resultados utilizando la prueba ENLACE media superior. *El trimestre Económico*. 79: 783-811.
- Fakhri A. S. 2020. Students' performance evaluation using machine learning algorithms. *College of Basic Education Researches Journal*. 16: 977-986. Doi: [10.33899/berj.2020.166006](https://doi.org/10.33899/berj.2020.166006)
- Fernández, A. T. 2003. Determinantes Sociales, Organizacionales, e Institucionales de los Aprendizajes en la Educación Primaria en México: Un análisis de tres niveles (2001).

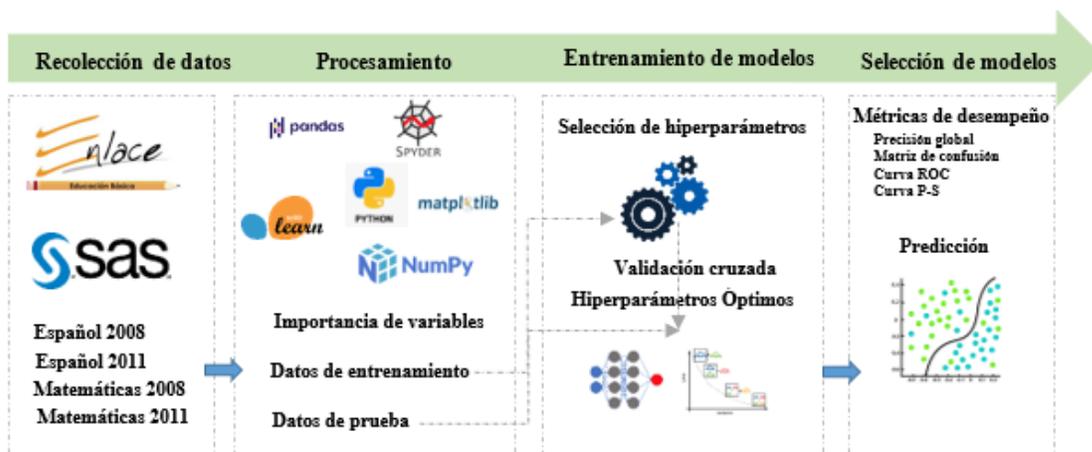
- Instituto Nacional para la Evaluación de la Educación. México. 153 p. <https://www.inee.edu.mx/wp-content/uploads/2019/01/P1C126.pdf>
- García J. J., J. Rodríguez S., y J. Torres. G. 2020. Influence of contextual variables on educational performance: a study using hierarchical segmentation trees. *Sustainability*. 12: 1-10. Doi: <https://doi.org/10.3390/su12239933>
- Géron A. 2019. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Segunda Edición. O'Reilly Media. 483 p.
- González-Camacho, J. M., G. De los Campos., P. Pérez., D. Gianola., J. E. Cairns., G. Mahuku., R. Babu., and J. Crossa. 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor Appl Genet*. 125: 759-771. Doi: <https://doi.org/10.1007/s00122-012-1868-9>
- Hussain, S., and M. Qasim. 2021. Student-Performer: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning. *Annals of Data Science*. <https://link.springer.com/article/10.1007%2Fs40745-021-00341-0>
- INEE (Instituto Nacional para la Evaluación de la Educación). 2019. Bases de datos INEE. Disponible en <https://www.inee.edu.mx/evaluaciones/bases-de-datos/>. (Consultado: enero 2021).
- INEGI (Instituto Nacional de Estadística y Geografía). 2020. Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Localidades. Disponible en <https://www.inegi.org.mx/app/ageeml/#>. (Consultado: noviembre 2020).
- Mandrekar, J. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*. 5, 1315-1316. Doi: <https://doi.org/10.1097/JTO.0b013e3181ec173d>
- Martínez, F., L. Contreras., E. González., J. Jornet., R. Martínez., F. Martínez., M. Pérez., F. Reyes., L. Santibañez., G. Solano., M. Tant., y A. Tristrán. 2015. Las pruebas ENLACE y EXCALE. Un estudio de validación. México: INEE. 119 p.
- Mayer, D., y E. E. Serván. 2009. Formación de la capacidad cognitiva en México: impactos económicos y de políticas públicas. *Estudios Económicos*. Número extraordinario: 83-122.
- Medina-Merino, R. F., y C. I. Ñique-Chacón. 2017. Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases*. 10: 165-189. Doi: <http://dx.doi.org/10.26439/interfases2017.n10.1775>
- Natekin A., and A. Knoll. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurobotics*. 7: 1-21. Doi: <https://doi.org/10.3389/fnbot.2013.00021>
- OCDE (Organización para la Cooperación y Desarrollo Económicos). 2005. *School Factors Related to Quality and Equity: Results from PISA 2000*. OECD Publishing, Paris. 157 p.

- Polyzou A., and G Karypis. 2018. Feature Extraction for Classifying Students Base don their Academic Performance. In: International Conference on Educational Data Mining. Proceedings of the 11th International Conference on Educational Data Mining. Boyer K. E., and M. Yudelson (eds.). Buffalo, NY. July, 2018. pp: 356-362.
- Quintana L. M., J. M. Flores A., S. Lazcano S., y V. M. Landassuri M. 2018. Ensamble de clasificadores para determinar el perfil académico del estudiante usando árboles de decisión y redes neuronales. *Research in Computing Science*. 147: 255-267.
- Rai, S., K. Shastry, S. Pratap, S. Kishore, P. Mishara, and H. Sanjaty. 2020. Machine Learning Aproach for Student Academic Performance Prediction. In: Evolution in Computational Intelligence. *Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020)*., Bhateja, V., M. Grobelnik, S. Peng, S. Sataphaty, and Y-D. Zhang (eds). Springer (ed.). September 2020. pp: 611-618.
- Raschka, S. 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. University of Wisconsin–Madison. Department of Statistics. arXiv repository. 49 p.
- Raschka, S., and V. Mirjalili. 2017. Python Machine Learning. Machine Learnin and Deep Learning with Python-scikit-learn, and TensorFlow. Packt Publishing. 622 p.
- Rogers, J., and S. Gunn. 2005. Identifyin Feature Relevance Using a Random Forest. In: Subspace, Latent, Structure and Feature Selection. Saunders, C., M. Grobelnik, S. Gunn, and J. Shawe-Taylor (eds). Bohinj, Slovenia. February 2005. School of Electronics and Computer Science. School of Electronics and Computer Science. pp: 173-184.
- Saito, T., and M. Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 10: 21. Doi: <https://doi.org/10.1371/journal.pone.0118432>
- Schonlau M., and R. Yuyan Z. 2020. The random forest algorithm for statistical learning. *The Stata Journal*. 20: 3-29. Doi: <https://doi.org/10.1177/1536867X20909688>
- SEP - IEIA. Secretaría de Educación Pública - Instituto de Evaluación e Ingeniería Avanzada, S.C. 2008. Enlace Educación básica. Manual técnico 2008. Disponible en [http://enlace.sep.gob.mx/ba/manuales\\_tecnicos/](http://enlace.sep.gob.mx/ba/manuales_tecnicos/). (Consultado: septiembre 2018)

## ANEXO

### DESCRIPCIÓN DE ESTRUCTURA DEL PROGRAMA EN PYTHON

En la colecta de datos, se obtuvieron 12 archivos en formato “.csv”, cuatro por entidad (ESP2008, ESP2011, MAT2008 y MAT2011). Se tienen 11 036 registros de Tlaxcala, 32 382 en Guerrero y 40 935 en Nuevo León. De los archivos, cada registro corresponde a un estudiante; y hay 13 variables predictoras (puntaje, características de contexto, ubicación geográfica de la escuela, etcétera). Estos archivos son la entrada para entrenar los modelos de aprendizaje automático. El procesamiento, análisis de los archivos e implementación de los modelos de aprendizaje automático se hizo en Python versión 3.8 con el entorno de desarrollo Spyder con las bibliotecas Numpy, Pandas, Matplotlib y Scikit-Learn. La primera etapa del procesamiento fue la preparación de los datos con dos actividades: en la primera, las variables categóricas  $t\_esc$ ,  $t\_loc$ ,  $n\_mar$ , se transforman a variables indicadores y se generan archivos de salida con 22 variables por entidad. En la segunda actividad, se ajusta el algoritmo RF y con  $CV = 5$  se obtienen los valores óptimos de los hiperparámetros. Con feature importance de Scikit-learn se tiene la importancia de las variables. La segunda etapa corresponde a la implementación de los algoritmos MLP y GB. El entrenamiento de estos modelos se hace en dos partes, en la primera, se seleccionan los hiperparámetros óptimos de los modelos a través de búsqueda en retícula con  $CV = 5$ . En la segunda, se hace  $CV = 5$  con el conjunto total de datos para obtener las diferentes métricas de desempeño. Las dos etapas anteriores se hacen para cada conjunto de datos de las tres entidades seleccionadas. El conjunto de datos de entrada se encuentra restringido debido a la privacidad de los datos, sin embargo, los archivos “.py” están disponibles a solicitud en el siguiente repositorio: [https://github.com/mmoralesh/Investigacion\\_ENLACE](https://github.com/mmoralesh/Investigacion_ENLACE) o al correo [mangel.moralesh@gmail.com](mailto:mangel.moralesh@gmail.com)



## RECOLECCIÓN DE DATOS Y PROCESAMIENTO

Para la implementación de los algoritmos se utilizaron diferentes bibliotecas de Python, principalmente de Scikit-Learn. De Raschka y Mirjalili (2017) se tomaron los códigos fuente y se adaptaron para aplicarlo a los datos de la prueba ENLACE.

```
#INVESTIGACIÓN-CLASIFICACIÓN DE NIVELES DE LOGRO PRUEBA ENLACE
#IMPORTACIÓN DEL DATASET
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

#IMPORTO DB POR ENTIDAD (EJEMPLO TLAXCALA-MATEMÁTICAS)
df=pd.read_csv('PENLACE_2011_MAT.csv')

#TRANSFORMO LAS VARIABLES EN DUMMYS
df=pd.get_dummies(df)
df.head()

#GUARDO DATASET (ejemplo Tlaxcala)
df.to_csv('mt1_2011.csv')

#SELECCIÓN DE CARACTERÍSTICAS DATOS ENLACE ESPAÑOL (ejemplo Tlaxcala)
df=pd.read_csv('et1_2008.csv')
df.columns = ['niv_esp', 'pun_mat', 'becario', 'turno', 'tip_sos', 'tip_loc', 'sexo', 'pob', 'lat', 'lon', 'alt', 'tip_esc_gen',
              'tip_esc_par', 'tip_esc_tec', 'tam_loc_100a249', 'tam_loc_15000ymas', 'tam_loc_2500a14999', 'tam_loc_250a499',
              'tam_loc_500a2499', 'tip_marg_alto', 'tip_marg_bajo', 'tip_marg_medio', 'tip_marg_nuyalto']

#SEPARAMOS EN MUESTRAS DE ENTRENAMIENTO Y PRUEBA (80-20)
np.random.seed(123) #Semilla
X = df.iloc[:, 1:].values
y = df.iloc[:, 0].values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, stratify=y, random_state=1)

#IMPORTANCIA DE LAS CARACTERÍSTICAS CON RANDOM FOREST
#CARACTERÍSTICAS EN ORDEN DE IMPORTANCIA
feat_labels = df.columns[1:]

forest1 = RandomForestClassifier(random_state=1,
                                criterion='gini',
                                max_depth=30,
                                min_samples_leaf=2,
                                min_samples_split=100,
                                n_estimators=300)

forest1.fit(X_train, y_train)
importances = forest1.feature_importances_
importances1=importances[0:6]

indices = np.argsort(importances)[::-1]
indices1=indices[0:6]

for f in range(X_train.shape[1]-13):
    print("%2d) %-s %f" % (f + 1, 30,
                          feat_labels[indices1[f]],
                          importances[indices1[f]]))
plt.bar(range(X_train.shape[1]-13), importances[indices1],
        align='center')

plt.xticks(range(X_train.shape[1]-13),
           feat_labels[indices1], rotation=90)
plt.xlim([-1, X_train.shape[1]-13])
plt.tight_layout()
plt.show()
```

## ENTRENAMIENTO Y SELECCIÓN DE MODELOS

```
# IMPLEMENTACIÓN DE ALGORITMOS
# Librerías de entrada
import pandas as pd
import numpy as np
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import HistGradientBoostingClassifier
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc

# Creación de listas para guardar resultados
np.random.seed(123) # Semilla aleatoria
names=[] # Guarda solo nombres de entidad_materia_año
best_par=[] # Hiperparámetros óptimos
score_M2=[] # Scores con CV
PG=[] # Precisión global
NC=[] # Muestras no clasificadas
RI=[] # Reporte de información
db=[] # Contiene las bases por entidad y año
# Guerrero
db.append(('MLP','HGB','gro_e08','egr_2008')) # Modelo1, modelo2, entidad_materia_año, materia_entidad_año
db.append(('MLP','HGB','gro_e11','egr_2011'))
db.append(('MLP','HGB','gro_m08','mgr_2008'))
db.append(('MLP','HGB','gro_m11','mgr_2011'))

# MLP-MultiLayerPerceptron-Classifier
def MLP():
    for a,b,nom, datab in db:
        # Ruta de archivos de entrada
        path = 'C:/Users/Miguel Angel/Desktop/articulos para tesis/tablas_vf/bd_tesis_mamh'
        all_files=(path + '/' + datab + '.csv')
        c=(a + '-' + nom)
        all_files2=(path + '/' + c + '/')
        # Creación de carpetas para guardar gráficas
        direct = os.path.join(path, c)
        os.mkdir(direct)
        # Lee archivo csv
        df = pd.read_csv(all_files)
        # Matriz de características y vector de clases
        X = df.iloc[:, 1:].values
        y = df.iloc[:, 0].values

        # Guarda nombre de entidades
        names.append(datab)

        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, stratify=y, random_state=2)
        #-----
        # Selección de hiperparámetros óptimos
        mlp_c = Pipeline([('scaler',StandardScaler()),
                        ('mlp',MLPClassifier(max_iter=200,
                                           random_state = 1))])

        param_grid = {
            'mlp_hidden_layer_sizes': [50,100,200, 250,300],
            'mlp_activation': ['tanh', 'relu', 'logistic'],
            'mlp_solver': ['sgd', 'adam', 'lbfgs'],
            'mlp_alpha': [0.0001,0.001, 0.1],
            'mlp_learning_rate': ['constant','adaptive']}

        gs = GridSearchCV(mlp_c,
                        param_grid,
                        scoring='accuracy',
                        n_jobs=-1,
                        cv=5)

        gs = gs.fit(X_train, y_train)
        best_par.append((nom,gs.best_params_))
        clf = gs.best_estimator_
        y_pred= clf.predict(X_test)
        #-----
        # Selección del mejor modelo/cálculo de PG
        cv_result = cross_validate(estimator=clf,
                                X=X,
                                y=Y,
                                cv=5,
                                n_jobs=1,
                                scoring='accuracy')

        cv_result1=cv_result['test_score']
        score_M1.append((nom,cv_result1))
        score_M2.append(cv_result1)
        P_G='MLPC Accuracy: %.3f +- %.3f' % (cv_result1.mean(), cv_result1.std())
        PG.append((nom,P_G))
        #-----
        # Muestras no clasificadas
        ncl='Misclassified samples: %d' % (y_test != y_pred).sum()
        NC.append((nom,ncl))
        print('Misclassified samples: %d' % (y_test != y_pred).sum())
```

```

#-----
#Matriz de confusión
confmat = confusion_matrix(y_true=y_test, y_pred=y_pred)
fig, ax = plt.subplots(figsize=(2.5, 2.5))
ax.matshow(confmat, cmap=plt.cm.twilight_shifted, alpha=0.3)
for i in range(confmat.shape[0]):
    for j in range(confmat.shape[1]):
        ax.text(x=j, y=i,
                s=confmat[i, j],
                va='center', ha='center')
plt.xlabel('Clases predichas')
plt.ylabel('Clases verdaderas')
plt.savefig(all_files2 + 'matrixC' + ".png")
plt.show()
#-----
#Gráfica Curva ROC
y_bin = label_binarize(y, classes=[0, 1, 2])
n_classes = y_bin.shape[1]
y_score = cross_val_predict(clf, X, y, cv=5, method='predict_proba')
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_bin[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
colors = cycle(['blue', 'red', 'green', 'pink'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i], color=color, lw=2,
             label='ROC, clase {0} (area = {1:0.2f})'
             ''.format(i, roc_auc[i]))
plt.plot([0, 1], [0, 1], 'k--', lw=2)
plt.xlim([-0.05, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('1-Especificidad')
plt.ylabel('Sensibilidad')
plt.legend(loc=(0, -.45), prop=dict(size=10))
plt.savefig(all_files2 + 'RocC' + ".png", bbox_inches='tight', dpi = 1000)
plt.show()
#-----
#Gráfica curva de precisión-sensibilidad
precision = dict()
recall = dict()
average_precision = dict()
for i in range(n_classes):
    precision[i], recall[i], _ = precision_recall_curve(y_bin[:, i], y_score[:, i])
    average_precision[i] = average_precision_score(y_bin[:, i], y_score[:, i])
precision["micro"], recall["micro"], _ = precision_recall_curve(y_bin.ravel(), y_score.ravel())
average_precision["micro"] = average_precision_score(y_bin, y_score, average="micro")
colors = cycle(['blue', 'red', 'green', 'pink'])

for i, color in zip(range(n_classes), colors):
    plt.plot(recall[i], precision[i], color=color, lw=2,
             label='P-S, clase {0} (area = {1:0.2f})'
             ''.format(i, average_precision[i]))
plt.xlim([-0.05, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('Sensibilidad')
plt.ylabel('Precisión')
plt.legend(loc=(0, -.45), prop=dict(size=10))
plt.savefig(all_files2 + 'S-E_C' + ".png", bbox_inches='tight', dpi = 1000)
plt.show()

```

MLP()