Análisis de la Influencia de la Vivienda y los Bienes Materiales en el Nivel Educativo en Jalisco, México, Mediante un Modelo de Inteligencia Artificial

[a]Hernández Martínez Jorge Iván

[a]Departamento de Computación, CINVESTAV, Instituto Politécnico Nacional.
[b]jivan.hernandez@cinvestav.mx

Abstract

La deserción escolar es un problema social y educativo en México y América Latina. Este estudio utiliza diferentes fuentes para examinar los factores asociados con la deserción escolar, incluyendo el entorno social, geográfico y familiar. Se citan estadísticas del Instituto Nacional de Estadística y Geografía (INEGI) para mostrar las tasas de deserción escolar en diferentes niveles educativos y estados de México. Además, se revisan estudios previos para examinar los factores que contribuyen a la deserción escolar, tales como el bullying, la delincuencia, la pobreza, y la calidad de la educación. Este estudio también presenta un modelo de clasificación para predecir el nivel de estudio de una persona utilizando diferentes características socioeconómicas y geográficas. Los resultados sugieren que el entorno social y geográfico de una persona es un factor importante en su educación, y destacan la necesidad de políticas públicas que aborden la brecha educativa y social existente en diferentes comunidades. Además, se utiliza la herramienta MLflow para la implementación y seguimiento de los modelos de clasificación. En conjunto, estos hallazgos pueden contribuir a mejorar las oportunidades educativas y sociales de la población en general.

I. Introducción

La deserción escolar es uno de los problemas más graves que enfrenta el sistema educativo mexicano. Si bien, ha venido disminuyendo con los años, según datos del Instituto Nacional de Estadística y Geografía (INEGI), la tasa de deserción escolar en México es del 2.94 % en el nivel de educación secundaria y del 12.75 % en educación media superior para el ciclo escolar 2020-2021 (de Estadística y Geografía (INEGI), 2021). Estas cifras son alarmantes, ya que la educación es uno de los pilares fundamentales para el desarrollo social y económico de cualquier país.

Existen diversas causas que pueden llevar a un estudiante a abandonar la escuela. Entre las más comunes se encuentran la falta de recursos económicos para continuar estudiando (Marcela, 2013), la falta de interés o motivación por parte del estudiante (S, 2012), el bullying o acoso escolar (Ruiz-Ramirez et al., 2016), problemas familiares o personales, la necesidad de trabajar para contribuir al ingreso familiar(Marcela, 2013), entre otros.

La deserción escolar no solo afecta al estudiante que abandona sus estudios, sino que también tiene consecuencias negativas para la sociedad en general. Los jóvenes que abandonan la escuela tienen mayores dificultades para acceder a empleos bien remunerados y, por lo tanto, tienen menos oportunidades de mejorar su calidad de vida (Ruiz-Ramirez et al., 2014). Además, la falta de educación puede aumentar la probabilidad de que estas personas caigan en la delincuencia o en situaciones de pobreza extrema (Millán & Pérez-Archundia, 2019).

Es fundamental que el gobierno y las instituciones educativas tomen medidas para reducir la tasa de deserción escolar en México. Esto implica no solo proporcionar recursos económicos para garantizar que los estudiantes puedan continuar sus estudios, sino también trabajar en la prevención y el tratamiento de problemas como el acoso escolar y los problemas familiares o personales que pueden llevar a los jóvenes a aban-

donar la escuela. La educación es un derecho humano fundamental y es responsabilidad de todos asegurarnos de que se garantice a todos los niños y jóvenes mexicanos el acceso a una educación de calidad.

En este trabajo, se llevará a cabo un estudio específico sobre la relación que existe entre aspectos de vivienda y bienes materiales con el nivel máximo de estudio alcanzado en la población de Jalisco, México. Para ello, se utilizará un modelo de inteligencia artificial que permitirá analizar una gran cantidad de datos del Censo de Población y Vivienda del año 2020 proporcionados por el INEGI. Este modelo permitirá identificar patrones y tendencias en los datos, lo que permitirá obtener información valiosa sobre la relación entre el nivel educativo y la calidad de vida en la población de Jalisco. Además, el uso de un modelo de inteligencia artificial permitirá una mayor precisión y rapidez en el análisis de los datos, lo que facilitará la toma de decisiones para mejorar la calidad de vida y la educación en la población de Jalisco, México.

A. Metodología

Explicación aqui del pipeline del proyecto

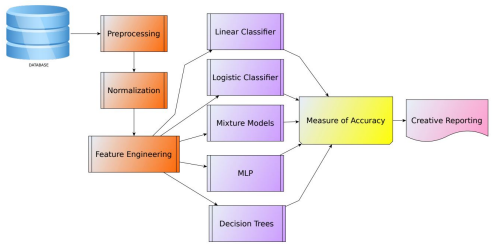

Figure 1: pipeline

A.1 Limpieza manual de Base de Datos

El INEGI, entidad autónoma del gobierno, llevó a cabo el Censo de Población y Vivienda 2020 en el estado de Jalisco con el objetivo de generar datos relacionados con la cantidad, composición y distribución geográfica de la población, así como también sus características socioeconómicas y culturales más relevantes. Además, el censo también permitió recolectar información sobre las viviendas, incluyendo detalles sobre materiales de construcción, servicios, equipamiento e instalaciones, entre otros aspectos.

La base de datos consta de 222 Features que cubren los aspectos antes mencionados, así como 108,041 muestras. Pero por la magnitud de una encuentra como esta, se tienen datos vacíos, que tienen que ser removidos ya que los modelos de machine Learning no pueden procesar y pueden afectar el rendimiento del modelo y la precisión de las predicciones al tener dificultades para entender los patrones subyacentes en los datos.

También es necesario hacer una limpieza manualmente de Features que para este estudio podrían no ser relevantes. Para este análisis se decidió tomar los features que describieran mejor un panorama mas general de las muestras, por ejemplo en la Figura 2 se muestra la decisión de tomar features generales y descartar los que tienen una relación directa sobre el mismo feature.

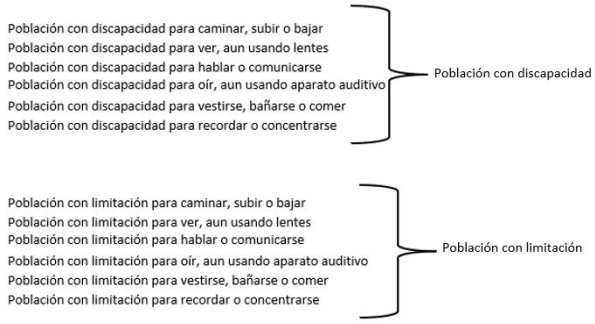

Figure 2: pipeline

Tambien se eliminaron Features como los mostrados en la Tabla con una dependencia total mutuamente excluyente, esto significa que de cumplirse uno de ellos, el otro por lo tanto no se cumple.

Features Iniciales Mutuamente excluyentes	
a	Viviendas particulares habitadas que disponen de energía eléctrica
a	Viviendas particulares habitadas que no disponen de energía eléctrica
b	Viviendas particulares habitadas que disponen de drenaje
b	Viviendas particulares habitadas que no disponen de drenaje
c	Viviendas particulares habitadas que disponen de agua entubada en el ámbito de la vivienda
c	Viviendas particulares habitadas que no disponen de agua entubada en el ámbito de la vivienda

La base de datos con los features utilizados

se puede obtener de (Poner link de github)

A.2 Preprocesamiento base de datos

Una vez se eliminaron cuidadosamente las columnas que pudieran no aportar nada al modelo, se procedió con el preprocesamiento de los datos para eliminar datos faltantes o nulos.Pero primero fue necesario considerar que si se eliminaban las muestras que tienen datos faltantes, el tamaño de la base de datos disminuiría a 9,316, lo que equivale a menos del 10% del tamaño total inicial. Como resultado, cualquier predicción basada en este conjunto de datos podría estar sesgada hacia un subgrupo específico de la población y no representar de manera adecuada a la muestra total de la población.
Para mitigar este problema primeramente se contabilizaron los datos nulos por columna, obteniéndose el gráfico de barras de la Figura 3

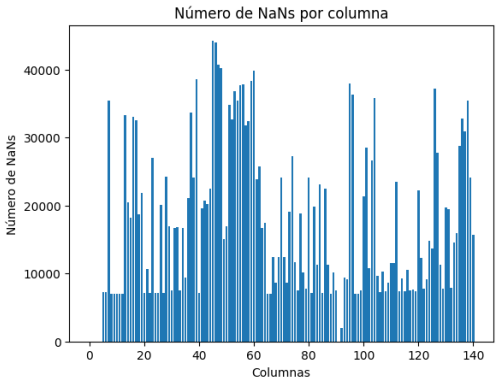

Figure 3: pipeline

En La Figura 3 es posible observar que hay columnas con mas de 30 mil muestras con datos nulos. Con la finalidad de conservar el mayor numero de features, pero sin comprometer la cantidad de muestras,se procedió a eliminar las columnas que tuvieran mas de 15 mil datos nulos. Con esto se logró conservar 63 de los 141 features iniciales para un total de 50,440 muestras.

A.3 Balanceo de datos

El modelo buscará los elementos y características que hacen que una persona pueda o no llegar a tener estudios posbasicos. La clasificación será binaria considerando a la población mayor a 15 años. Considerando a las personas analfabetas, sin estudios, primaria sin terminar, primaria terminada, secundaria incompleta y secundaria terminada dentro de una sola categoría. Mientras que la población con algún grado igual o mayor a posbasica (preparatoria, universidad, posgrado) serán considerados en otra categoría.
Con las etiquetas propuesta en este trabajo, la segmentación de ambos grupos se muestra en la tabla 1.

Personas mayores a 15 años con educación menor a posbasica	Personas mayores a 15 años con educación mayor a posbasica
13970	36470

Table 1: Etiquetas

Se puede observar una diferencia entre las cantidades de datos disponibles para cada categoría. Utilizar estos como datos de entrenamiento con una gran disparidad puede tener efectos negativos al momento de esperar que el modelo aprenda a reconocer patrones de todas las clases de manera justa. Si los datos no están balanceados, el modelo puede tener un sesgo hacia las clases sobre-representadas, lo que afecta su precisión en las clases sub-representadas y su capacidad de generalización. Por lo tanto es importante equilibrar el conjunto de datos de entrenamiento para garantizar que el modelo sea justo y preciso para todas las clases y pueda generalizar bien.
Para reducir la disparidad sin tener que reducir el numero de datos, se utilizó un algoritmo de sobremuestreo llamado SMOTE (Synthetic Minority Over-sampling Technique), con la implementación de este algoritmo se aumentó el tamaño de la clase minoritaria sintetizando nuevas instancias a partir de las instancias existentes. En lugar de crear copias exactas de las instancias minoritarias, se crearon nuevas instancias sintéticas interpolando entre las instancias minoritarias cercanas. De esta forma se logró duplicar la cantidad de datos que se tenían para la categoría de personas con educación

menor a posbasica. La nueva distribución de las etiquetas se observa en la Tablas 2

Personas mayores a 15 años con educación menor a posbasica	Personas mayores a 15 años con educación mayor a posbasica
27940	36470

Table 2: Etiquetas después de algoritmo SMOTE

Después se validó que los promedios y desviaciones estándar de los nuevos datos sintéticos, fuera similar a los datos iniciales. En la Tabla 3 se pueden observar la comparativa de media y desviación estándar de las 5 primeros características

	Feature	MUN	LOC	AGEB	MZA	POBTOT
-2*Datos iniciales	Media	78.28289191	74.86936292	1256.82269148	18.1392985	308.2765927
	STD	33.61213565	214.32635899	1524.92457022	25.77896093	2735.20832111
-2*Datos sintéticos	Media	78.70773154	76.25241042	1252.73833211	17.64448315	284.8669772
	STD	32.59911265	216.86024899	1517.38324897	19.16892677	1703.91935758

Table 3: Media y desviación estándar entre las primeras 5 características (features) y los datos sintéticos generados con SMOTE

Para tener clases aun mas balanceadas de lo que se obtuvo con el algoritmo SMOTE, se utilizó una reducción por submuestreo aleatorio sobre la clase 1, quedando así dos clases completamente balanceadas con 27940 datos cada una.

A.4 Selección de mejores Características

Se procedió a encontrar las mejores características que describieran mejor el modelo, reduciendo así la dimencionalidad del conjunto de datos. Para ello se implementó un algoritmo de backward selection, donde a través de generar grupos de features en todas las posibles combinatorias posibles, se calcula la media de cada feature sobre todos los datos y luego se busca el grupo que maximice la norma o separación entre clases (Figura 4). Visto desde el espació de los features, es buscar el grupo de features tales que separen mejor las dos categorías.
Para este trabajo se utilizó backward selection para encontrar el grupo de los mejores features que capaces de generar mejores predicciones. El numero de features no se especifica ya que se dejó como hyperparametro del propio modelo.

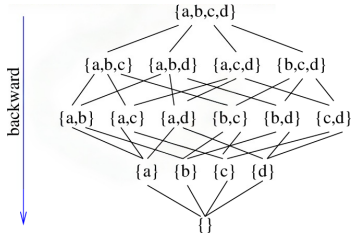

Figure 4: Algoritmo Backward selection

Después de haber obtenido los datos mediante el algoritmo de backward selection, se procedió a dividirlos en tres grupos: datos de entrenamiento, datos de validación y datos de prueba, en una proporción de 70%, 20% y 10% como se muestra en la Figura 5, respectivamente. Esto permitirá evaluar de manera más precisa el rendimiento de los modelos en datos que no fueron utilizados durante el entrenamiento.

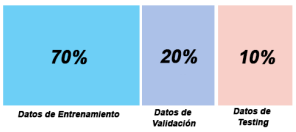

Figure 5: División del Dataset en conjunto de datos de entrenamiento, validación y testing con los que serán entrenados los modelos y se obtendrán sus métricas

A.5 Modelos de Clasificación

Los modelos de clasificación utilizados así como los hiperparametros (HP) estudiados fueron los siguientes:

- Mixture of Gaussians (MG)
  - HP: Numero de pasos
  - HP: Numero de Features
- Multilayer perceptron (MLP)
  - HP: Topology
  - HP: Steps
  - HP: Learning rate

  - HP: Threshold
  - HP: Numero de Features
- Lineal Clasifier(LIC)
  - HP: Steps
  - HP: learning rate
  - HP: Numero de Features
- Logistic Clasifier (LOGC)
  - HP: Steps
  - HP: learning rate
  - HP: Lamda
  - HP: Threshold
  - HP: Numero de Features
- Decision Tree (DT)
  - HP: Max depth
  - HP: Min Gini Threshold
  - HP: Numero de Features
- AdaBoost (AB)
  - HP: Numero de Stumps
  - HP: Número de Features

Las métricas estudiadas en todos ellos fueron Presicion, Recall y Accuraccy, y el caso especifico del MLP También se estudio el comportamiento de Loss.
Para el control de hiperparametros y estudio de las diferentes métricas obtenidas, se utilizó la Herramienta de MlFlow (Project, 2021)

II. Resultados

Cada modelo utilizado en el análisis de datos tiene ciertas ventajas y limitaciones, y algunos pueden no ser efectivos para ciertos tipos de datos. Por esta razón, se recomienda utilizar diversos modelos y estudiar las predicciones de cada uno de ellos para entender mejor el conjunto de datos y hacer mejores predicciones. Cada modelo cuenta con ciertos hiperparámetros que pueden ser ajustados para mejorar su rendimiento. A continuación, se presentan los resultados obtenidos con cada modelo utilizado en el estudio.

A. Backward Selection

Aunque este modelo no se usó como clasificador, se utilizó para identificar las características más relevantes del conjunto de datos. Estas características se identificaron en el espacio vectorial de los features, y se definieron como aquellas cuyos centros de masa estaban más separados entre sí. En la figura 6 se pueden observar los datos con los tres atributos más destacados mostrados en la tabla 4, los cuales lograron aumentar significativamente la separación entre los datos en el espacio de features.

Identificador Feature	Significado
GRAPROES_F	Grado promedio de escolaridad de la población femenina
GRAPROES_M	Grado promedio de escolaridad de la población masculina
PRO_OCUP_C	Promedio de ocupantes por cuarto en viviendas particulares habitadas

Table 4: Tres mejores Features obtenidos con algoritmo de Backward Selection

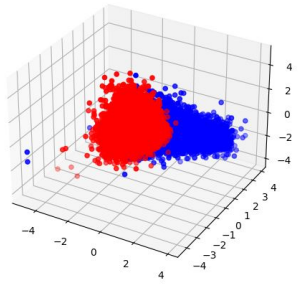

Figure 6: Datos vistos en el espacio de los 3 features:“GRAPROES_F” ,“GRAPROES_M”, “PRO_OCUP_C”

Este resultado es sumamente interesante, ya que demuestra que existen diferencias significativas entre las personas que tienen estudios superiores a secundaria y aquellas que no, incluso antes de aplicar un modelo de predicción. En particular, se encontró que la diferencia radica en el promedio de grado obtenido por las personas en su entorno cercano, así como el promedio de personas que comparten una habitación en viviendas particulares habitadas. Esto sugiere que una persona que vive en una comunidad más educada tiene más probabilidades de continuar sus estudios. No obstante, este parámetro está estrechamente relacionado con el número de personas que comparten una

habitación en una casa, lo que a su vez puede ser un indicador del poder adquisitivo del hogar. En general, se asocia un menor número de personas por habitación con un mayor poder adquisitivo en el hogar.

Un resultado similar se obtiene al encontrar los 9 mejores features que clasifican mejor el grado que puede llegar a tener una persona, siendo los mostrados en la tabla 4 mas los mostrados en la Tabla

Identificador Feature	Significado
PROM_OCUP	Promedio de ocupantes en viviendas particulares habitadas
PROM_HNV	Promedio de hijas e hijos nacidos vivos
AGEB	Clave del AGEB
LOC	Clave de localidad
REL_H_M	Relación hombres-mujeres
MUN	Clave de municipio o demarcación territorial

Table 5: Mejores Features ordenados de mayor relevancia a menor (decendente) para la clasificacion de personas por su grado de estudios

Los resultados presentados en la Tabla 5 muestran la relevancia del número de hijos que tiene una persona, así como la relación entre hombres y mujeres en una comunidad, en el ámbito educativo. Además, se encontraron características relacionadas con la localidad donde reside una persona, lo que sugiere una brecha educativa entre las comunidades más pequeñas y aisladas de las grandes urbes y ciudades, donde se concentran la mayoría de las escuelas y centros de estudio. Es interesante observar cómo estos factores pueden afectar significativamente la educación de una persona y resaltan la importancia de considerar no solo los aspectos individuales, sino también el entorno social y geográfico en el que se desenvuelve.

B. Mixture of Gaussians

Para el modelo de Mixture of Gaussians utilizado en este trabajo, se exploró un rango de 2 a 5 en el número de pasos necesarios para ajustar el modelo, así como un rango de 3 a 8 características (features) que se utilizaron para realizar la predicción. Esta búsqueda de hiperparámetros permitió determinar la mejor combinación de pasos y características para obtener un modelo óptimo y preciso en la predicción de los datos estudiados.

Los resultados de las diferentes corridas se pueden observar en la Figura 7

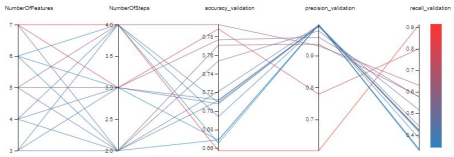

Figure 7: Grafico de Coordinadas Paralelas para el modelo de Mixture of Gaussian

Los mejores resultados se obtuvieron con los hyperparametros 4 y 5 para el numero de pasos y el numero de características respectivamente, dando así las métricas mostradas en la tabla 6 para los datos de entrenamiento así como de validación.

Métricas	Resultado
Accuracy Entrenamiento	0.795
Accuracy Validación	0.793
Precisión Entrenamiento	0.939
Precision Validación	0.93
Recall Entrenamiento	0.634
Recall Validación	0.629

Table 6: Metricas de mejor corrida en Mixture of Gaussians

Dentro de las métricas obtenidas, se priorizo buscar aquella con el mejor accuracy ya que esta mide la mide la proporción de predicciones correctas sobre el total de predicciones. Es útil cuando los costos de los falsos positivos y los falsos negativos son similares y no hay un sesgo significativo en los datos como en este caso.

En la figura -y- se muestras los gráficos en 2 y 3 dimenciones para la mejor corrida donde se puede observar representado una proyección del hyperespacio de 4 dimenciones (4 features) a un espacio de 2 y tres dimenciones para su análisis.

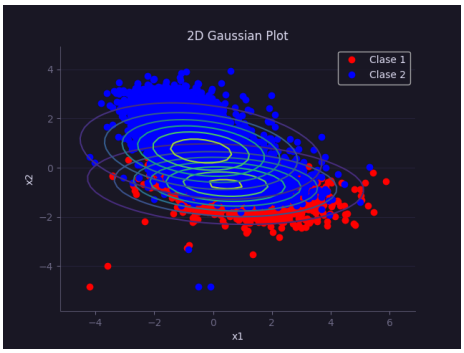

Figure 8: Grafico 2D de ajuste usando MG

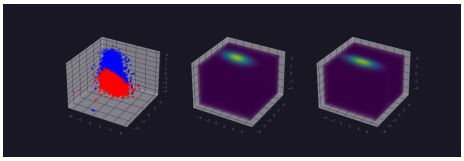

Figure 9: Gráfico 3D de ajuste usando MG. El primer gráfico corresponde a los datos representados en 3d, El segundo gráfico representa el volumen generado por la gaussiana que encierra a los datos de personas con estudios menor a secundaria. El tercer gráfico (derecha), representa el volumen generado por la gaussiana que encierra a los datos de personas con estudios mayor a secundaria

En las Figuras 8 y 9 se puede observar que los datos no presentan una separación perfecta, por lo que su clasificación no puede ser perfecta.

C. Multilayer perceptron

Para el Modelo de Multilayer Perceptron (MLP) se exploró un rango de 3 a 30 features, de 1000 a 10000 steps en pasos de 1000, y un learning rate de 0.005 a 0.05 en pasos de 0.005.

Las diferentes combinaciones de hyperparametros obtenida se observa en el grafico de cordenadas en la Figura

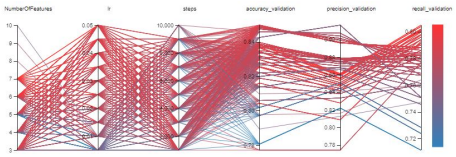

Figure 10: Gráfico de Coordinadas Paralelas para el modelo de MLP.

Los mejores resultados se obtuvieron para los hyperparametros mostrados en la tabla 8 resultando así en las metricas de la Tabla 8

Hyperparametro	Valor
Numero de Features	3
Learning Rate	0.05
Steps	9000
Topologia	[3,20,15,2]
threshold	0.0005

Table 7: Métricas de mejor corrida en MLP

Métricas	Resultado
Accuracy Entrenamiento	0.854
Accuracy Validación	0.85
Precisión Entrenamiento	0.904
Precision Validación	0.896
Recall Entrenamiento	0.793
Recall Validación	0.788

Table 8: Mejores Métricas obtenidas en MLP

En la Figura se observa como fue el comportamiento del loss con la configuración de los mejores hyperparametros, donde podemos observar que el mínimo valor lo alcanza cerca de los primeros 15 pasos.

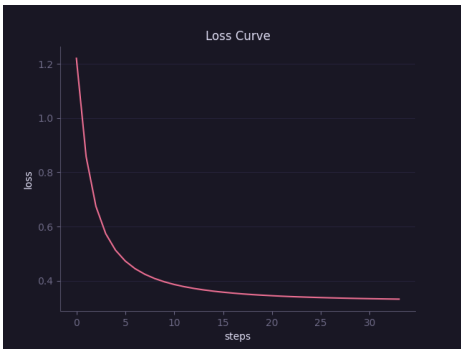

Figure 11: Gráfico de Coordinadas Paralelas para el modelo de MLP.

En la Figura 12se observa la relación entre las métricas precision y recall, donde se observa una relación inversamente lineal lo que significa que si la precisión aumenta mientras que el recall disminuye, el modelo está haciendo menos predicciones positivas, pero es más probable que las que haga sean correctas. Es decir, el modelo se está volviendo más selectivo y está eliminando las predicciones que tienen una alta probabilidad de ser falsas positivas.

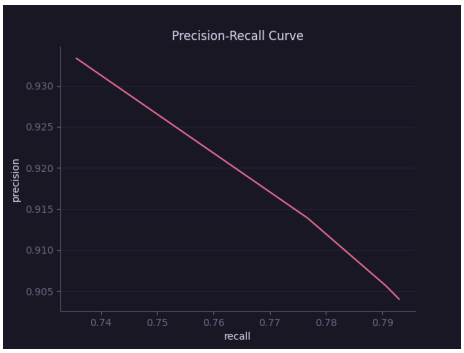

Figure 12: Gráfico Precisión-Recall MLP

En la Figura 13 se observa la matriz de confusión para el mejor modelo de mlp, donde se puede observar una clara mayoría en la diagonal de predicciones correctas, pero donde claramente tenemos elementos que fueron clasificados erróneamente. Específicamente el generó muchas predicciones falsas positivas, lo que conlleva a obtener un recall mas bajo.

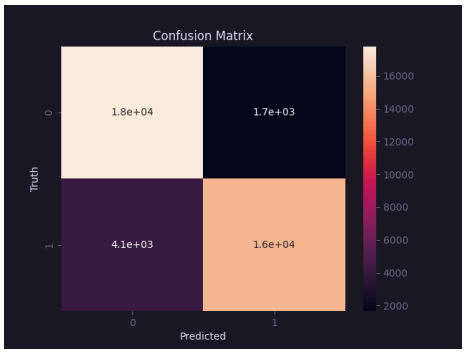

Figure 13: Matrix de confusion MLP.

D. Lineal Clasifier

Para este trabajo se utilizaron dos tipos de modelos lineales: Descenso del gradiente y Ridge Regresión.

D.1 Descenso del Gradiente

El rango de búsqueda en los hyperparametros fue de 3 a 10 features, de 1000 a 10000 en pasos de 1000 para el máximo numero de steps y de $4.4E^{-7}$ a $4.4E^{-6}$ en pasos de $1E^{-7}$ para el learning rate.

Las diferentes combinaciones de hyperparametros se observa en el gráfico de coordenadas en la Figura 14

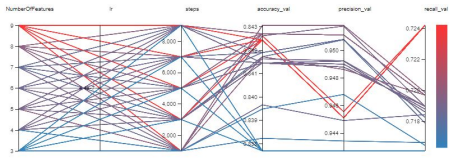

Figure 14: Gráfico de Coordinadas Paralelas para el modelo de Gradient descent

Los mejores resultados se obtuvieron para los hyperparametros de mostrados en la Tabla 9, resultando así en las métricas de la Tabla 10

Hyperparametro	Valor
Numero de Features	8
Learning Rate	$4.4E^{-7}$
Steps	1000

Table 9: Hyperparametros de mejor corrida en Gradient descent

Métricas	Resultado
Accuracy Entrenamiento	0.842
Accuracy Validación	0.843
Precisión Entrenamiento	0.955
Precision Validación	0.952
Recall Entrenamiento	0.721
Recall Validación	0.720

Table 10: Mejores Métricas obtenidas en Gradient descent

D.2 Ridge Regression

El rango de búsqueda en los hyperparametros fue de 3 a 10 features y de 0 a 100 en pasos de 1 para el coeficiente de regularización lambda.

Las diferentes combinaciones de hyperparametros se observa en el gráfico de coordenadas en la Figura 15

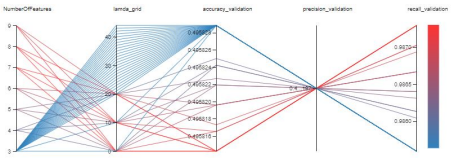

Figure 15: Gráfico de Coordinadas Paralelas para el modelo de Ridge Regression

Los mejores resultados se obtuvieron para los hyperparametros de mostrados en la Tabla 11, resultando así en las métricas de la Tabla 12

Hyperparametro	Valor
Numero de Features	3
Lambda	20

Table 11: Hyperparametros de mejor corrida en Ridge Regression

Métricas	Resultado
Accuracy Entrenamiento	0.517
Accuracy Validación	0.496
Precisión Entrenamiento	0.51
Precision Validación	0.496
Recall Entrenamiento	0.1
Recall Validación	0.986

Table 12: Mejores Métricas obtenidas en Ridge Regression

E. Logistic Clasifier

El rango de búsqueda en los hyperparametros fue de 3 a 10 features, de 500 a 2000 en pasos de 200 para el maximo numero de steps, de 0 a 10 en pasos de 1 para el coeficiente de regularización lambda y de $2.3E^{-3}$ a $2.3E^{-2}$ en pasos de $1E^{-3}$ para el parámetro de Learning Rate.

Las diferentes combinaciones de hyperparametros se observa en el gráfico de coordenadas en la Figura 16

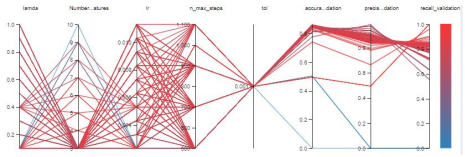

Figure 16: Gráfico de Coordinadas Paralelas para el modelo de logistic Regression

Los mejores resultados se obtuvieron para los hyperparametros de mostrados en la Tabla 13, resultando así en las métricas de la Tabla 14

Hyperparametro	Valor
Numero de Features	9
Learning Rate	$5.3E^{-3}$
Steps	500
Lambda	0.1

Table 13: Hyperparametros de mejor corrida en Logistic Regression

Métricas	Resultado
Accuracy Entrenamiento	0.857
Accuracy Validación	0.863
Precisión Entrenamiento	0.927
Precision Validación	0.892
Recall Entrenamiento	0.776
Recall Validación	0.824

Table 14: Mejores Métricas obtenidas en logistic Regression

F. Decision Tree

El rango de búsqueda en los hyperparametros fue de 3 a 10 features, de 2 a 7 en pasos de 1 para la profundidad del modelo y de 0.1 a 0.25 en pasos de 0.05 para el threshold de Gini en alguno de las ramas del árbol.

Las diferentes combinaciones de hyperparametros se observa en el gráfico de coordenadas en la Figura 18

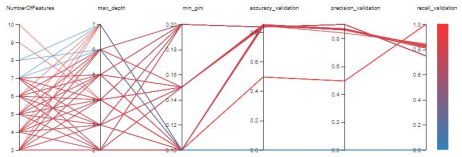

Figure 17: Gráfico de Coordinadas Paralelas para el modelo de Decision Tree

Los mejores resultados se obtuvieron para los hyperparametros de mostrados en la Tabla 13, resultando así en las métricas de la Tabla 16

Hyperparametro	Valor
Numero de Features	3
Profundidad Máxima	6
Gini Threshold	0.1

Table 15: Hyperparametros de mejor corrida en Modelo de Decision Tree

Métricas	Resultado
Accuracy Entrenamiento	0.854
Accuracy Validación	0.853
Precisión Entrenamiento	0.873
Precision Validación	0.869
Recall Entrenamiento	0.831
Recall Validación	0.829

Table 16: Mejores Métricas obtenidas en modelo de Decision Tree

En la Figura se muestra el diagrama de árbol para la corrida con los hyperparametros de la Tabla 15.

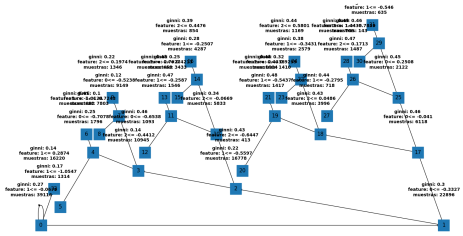

Figure 18: Gráfico de árbol para modelo entrenado con los hyperparametros de Tabla 15

G. AdaBoost

El rango de búsqueda en los hyperparametros fue de 3 a 10 features y 2 a 7 para el numero de stumps (también conocidos como weak learners).

Las diferentes combinaciones de hyperparametros se observa en el gráfico de coordenadas en la Figura 19

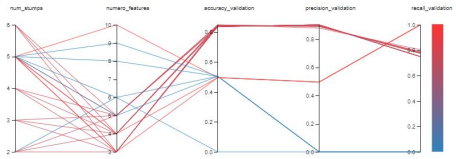

Figure 19: Gráfico de Coordinadas Paralelas para el modelo de AdaBoost

Los mejores resultados se obtuvieron para los hyperparametros de mostrados en la Tabla

13, resultando así en las métricas de la Tabla 18

Hyperparametro	Valor
Numero de Features	6
Numero de Stumps	5

Table 17: Hyperparametros de mejor corrida en Modelo de AdaBoost

Métricas	Resultado
Accuracy Entrenamiento	0.848
Accuracy Validación	0.848
Precisión Entrenamiento	0.886
Precision Validación	0.879
Recall Entrenamiento	0.801
Recall Validación	0.804

Table 18: Mejores Métricas obtenidas en modelo de AdaBoost

III. Discusión y conclusiones

Los resultados obtenidos de los diferentes modelos de clasificación utilizados demuestran que existen características significativas en los datos que permiten predecir el nivel de estudio de una persona. Además, la identificación de los atributos más relevantes destaca la importancia del entorno social y geográfico en el que una persona se desenvuelve para su educación. En particular, se encontró que el promedio de grado obtenido por las personas en su entorno cercano y el promedio de personas que comparten una habitación en viviendas particulares habitadas son factores determinantes. Asimismo, el número de hijos que tiene una persona, la relación entre hombres y mujeres en una comunidad y la localidad donde reside también pueden afectar significativamente su educación. Estos hallazgos sugieren la necesidad de políticas públicas que aborden la brecha educativa y social existente en diferentes comunidades, lo que puede mejorar las oportunidades educativas de la población en general.

References

de Estadística y Geografía (INEGI), I. N. (2021). Tasa de deserción escolar en educación básica, media superior y superior, por entidad federativa. https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=9171df60-8e9e-4417-932e-9b80593216ee. (Accedido el 30 de marzo de 2023)

Marcela, R. (2013, 01). Factores asociados al abandono y la deserción escolar en américa latina: Una mirada de conjunto. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, *11*, 33-59.

Millán, H., & Pérez-Archundia, E. (2019, 01). Educación, pobreza y delincuencia: ¿nexos de la violencia en méxico? *Convergencia Revista de Ciencias Sociales*, *26*, 1. doi: 10.29101/crcs.v26i80.10872

Project, M. (2021). *Mlflow.* https://mlflow.org/. (Accessed: Abril 17, 2023)

Ruiz-Ramirez, R., Garcia Cué, J., & Olvera, M. (2014, 07). Causas y consecuencias de la deserción escolar en el bachillerato: caso universidad autónoma de sinaloa. *Ra Ximhai*, *10*, 51-74. doi: 10.35197/rx.10.03.e1.2014.04.rr

Ruiz-Ramirez, R., Zapata Martelo, E., Garcia Cué, J., Olvera, M., Corona, B., & Martínez, G. (2016, 06). Bullying en una universidad agrícola del estado de méxico. *Ra Ximhai*, 105-126. doi: 10.35197/rx.12.01.2016.06.rr

S, M. (2012). *Causas, consecuencias y prevención de la deserción escolar: Un manual de auto ayuda para padres, maestros y tutores.* Palibrio. Retrieved from https://books.google.com.mx/books?id=8CsqYvnFFL0C