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

This work presented a detailed review of Density Functional Theory (DFT) and the Davidson method used in DFT as a k-eigenvalue solver. This method is widely used in the most popular DFT software in the solving of the self-consistency of iterative Kohn-Sham(KS) equations. This article explains the mathematics of this method and provides an easy-to-understand code.

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 abandonar 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

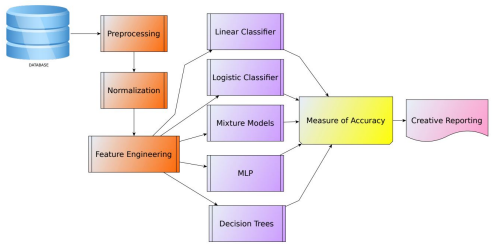

Figure 1: pipeline

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.

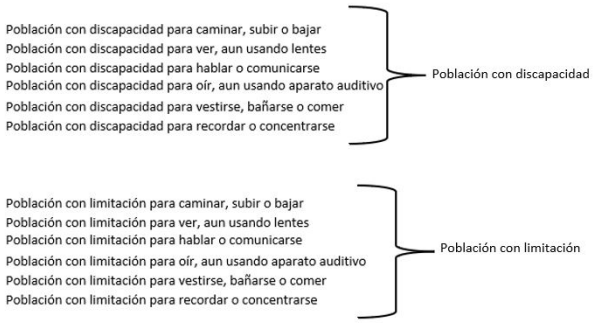

Figure 2: pipeline

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

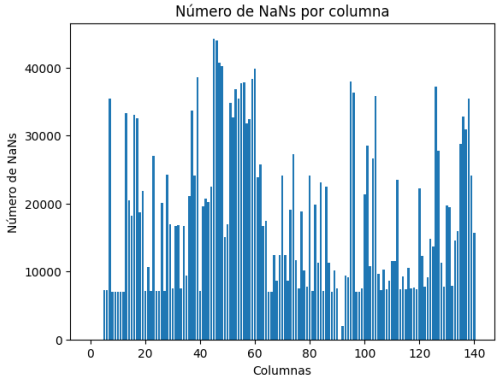

Figure 3: pipeline

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.

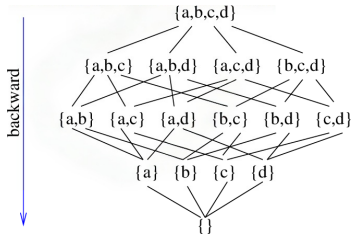

Figure 4: Algoritmo Backward selection

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
- Multilayer perceptron (MLP)
  - HP: Topology
  - HP: Steps
  - HP: Learning rate
  - HP: Threshold
- Lineal Clasifier(LIC)
  - HP: Steps
  - HP: learning rate
- Logistic Clasifier (LOGC)
  - HP: Steps
  - HP: learning rate
  - HP: Lamda
  - HP: Threshold

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 metricas obtenidas, se utilizó la Herramienta de MlFlow (Project, 2021)

A. Apéndice A

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