Análisis de datos de los estudiantes en la Educación Formal en Guatemala del año 2023

Analysis of student data in Formal Education in Guatemala in 2023

Resumen

OBJETIVO: encontrar información relacionada con los estudiantes para así conocer cuales son las necesidades de cada área, sector o departamento en el ámbito de la educación formal en Guatemala. MÉTODO: se realizará a partir de cuatro fases, i.) primero se obtendrán variables relacionadas entres si que puedan ser de alto valor para un análisis más profundo, esto usando métodos de minería de datos como la relación a priori o fpgrowth, ii.) en la segunda fase se hará una segunda evaluación con estas variables altamente relacionadas y se intentará predecir su valor dependiendo de la otra y de otras variables que se consideren importantes en el estudio, iii.) en la tercera fase se usarán métodos del random forest para hacer análisis y predicciones más robustas de la relación entre variables, iv.) y como última fase se hará un modelo de redes neuronales para predecir variables desde esta vía, así poder comparar modelos de predicción y resultados entre cada una de ellas. RESULTADOS: a) El resultado final de cada estudiante, el aprobar o el no hacerlo no tiene relación alta con ninguna de las variables, b) El pueblo de pertenencia esta altamente representado por mayas y ladinos, y una de las variables más estrechas con este mismo es el departamento, c) El sector, el área, plan estudiantil y jornada son variables que se complementan en sí, sobre todo la jornada si es doble con el plan estudiantil, d) Hay diferencias claras de clase y de sectores sociales dependiendo del departamento, el departamento es una variable que tiene mucho peso entre las demás, e) Cómo último punto, los niveles y grado generan mucho ruido en la información ya que tienen muchas veces relación entre sí, pero no relación en otra data importante. CONCLUSIONES: Las diferencias en niveles sociales y económicos se ven reflejados entre departamentos, haciendo que haya una brecha, no solo en acceso a la educación si no cómo los estudiantes se terminan relacionando con el área educacional del país, siendo que departamentos como Guatemala cuanta con una alta población ladina, y con una gran representación del sector privado, como otros departamentos centrales, mientras los departamentos más alejados al

centro tienen más representación en jornadas distintas a la matutina y plan diario, siendo un reflejo del posible trabajo infantil en las que estén los niños de estos lugares.

Palabras clave

Área, sector, jornada, plan de estudio, pueblo de pertenencia, redes neuronales, random forest, apriori, fpgrowth, minería de datos.

Abstract

OBJECTIVE: To gather information related to students to identify the needs of each area, sector, or department within the formal education system in Guatemala. METHOD: The process will be carried out in four phases: i) First, variables that are interrelated and potentially valuable for deeper analysis will be identified using data mining methods such as the Apriori algorithm or FP-Growth. ii) In the second phase, these highly related variables will be further evaluated, and their values will be predicted based on each other and other variables deemed important for the study. iii) The third phase will employ Random Forest methods to perform a more robust analysis and prediction of relationships between variables. iv) Finally, a neural network model will be developed to predict variables through this method, allowing for comparisons between prediction models and their respective results. RESULTS: a) The final outcome for each student, whether they pass or fail, does not have a strong correlation with any of the variables. b) Ethnic affiliation is predominantly represented by Mayans and Ladinos, and it is closely associated with the department of residence. c) Sector, area, educational program, and schedule are interrelated variables, particularly the double-shift schedule when combined with the educational program. d) There are clear class and social sector differences depending on the department, with the department being a highly influential variable. e) Levels and grades generate significant noise in the data because they are often related to each other but show little correlation with other important data. CONCLUSIONS: Social and economic disparities are reflected across departments, creating a gap not only in access to education but also in how students interact with the educational system in the country. For instance, departments like Guatemala have a high Ladino population and a significant representation from the private sector, similar to other central departments. In contrast, departments

farther from the center show higher representation in non-morning schedules and daily programs, potentially reflecting child labor in those areas.

Keywords

Área, sector, schedule, study plan, ethnic affiliation, neural networks, random forest, apriori, fpgrowth, data mining.

Introducción

La educación formal en Guatemala dependiendo del nivel al que se está hablando tiene un porcentaje más reducido, desde preprimaria se ve una tasa de asistencia de un 43% en preprimaria, un 79.2 en primaria y un 22.8 en secundaria baja y alta, siendo ya índices muy bajos hasta para la región de Latinoamerica (SITEAL y UNESCO, 2019), y se ve un decrecimiento en algunas de estos niveles a través de los años. No hay estudios tampoco muy actualizados, por lo que se necesita hacer avances en los análisis de esta información para poder mejorar y proporcionar áreas de oportunidad necesarias para la mejora del sistema educativo. Según la Unesco Guatemala es uno de los países con una de las tasas de deserción más altas de Latinoamérica y con mayor diferencia socioeconómica entre estudiantes, esto genera problemas estructurales en la capacidad de proveer educación desde centros que no tienen recursos al ser marginados.

El análisis de datos y minería de datos es esencial en tiempos actuales para encontrar relaciones entre variables que pueden no parecer relacionadas en un principio, en el caso de educación puede servir para encontrar y analizar estos problemas estructurales desde el punto de vista de departamentos, pueblo de pertenencia, sector, área y entre otras. Se usarán modelos de predicción y de asociación de variables para poder evaluar la relación que hay entre comunidades y población con la educación que reciben.

Materiales y métodos

La investigación se llevó a cabo entre el mes de noviembre y diciembre del 2024, donde se tomó la información del ine, de los datos de todo el año del 2023 empleando varias fases de minería y análisis de datos, el análisis se realizo en R en su mayoría, y en Python las redes neuronales.

En la primera fase se realizo un análisis de relación de variables en R, en donde se obtuvieron a través del algoritmo apriori y fpgrowht las variables más relacionadas, como también se fueron desechando las variables que más ruido hacían en el dataset de datos.

En una segunda fase con las variables que se encontraron una mayor relación se hicieron predicción a través de árboles de decisión y posteriormente usando random forest como modelo de predicción, se usó la herramienta de R, donde se obtuvo un contraste en las predicciones y efectividad en obtener resultados a partir de estos modelos.

En una tercera fase se realizó un par de predicciones usando redes neuronales simples para obtener que tan probable es que una condición se cumpla a partir de otras variables, en esta fase su harán transformaciones de datos y creación de variables que permitieran hacer el análisis porcentual entre 0 y 1.

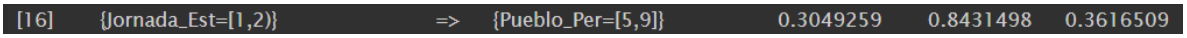
Resultados y discusión

Primero se usó el algoritmo apriori y después de varias pruebas se usó un soporte de 0.4 y un 0.5 de confianza, se usó esta configuración al no tener una relación alta entre variables.

```
[17]    {Nivel=[2,5], Pueblo_Per=[5,9]}    =>    {Repitente=[2,9]}
```

Encontramos relaciones poco fiables ya que se mira como al ser ladino está relacionado a no ser repitente, pero después veremos que el ser repitente no está altamente relacionado con el pueblo de pertenencia. En este primer proceso de la fase 1 se encontró que el Nivel hacía mucho ruido al tener una relación alta con el grado y poca relevancia con las otras variables.

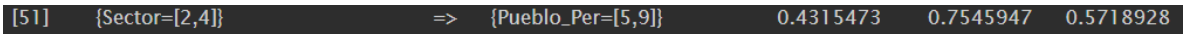
En una segunda corrida se decidió disminuir el soporte a 0.3, para tener relaciones más raras y no tan obvias como el que el grado 6 pertenece al nivel de primaria, con resultados más interesantes como



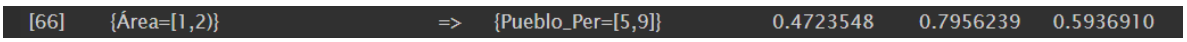
Relacionando la jornada estudiantil matutina, con el pueblo ladino.

Relacionando el área rural con la jornada estudiantil no matutina.

Relaciona el no ser del sector público y que el área de la persona sea urbana.



Relaciona que el pueblo ladino normalmente no estudia en el sector público



Relaciona que el estudiante de área urbana tiene una tendencia a pertenecer al pueblo ladino.

También en este proceso se encontró que las variables de Graduando y Repitente generan mucho ruido innecesario. Seguidamente se hizo un análisis con el algoritmo fpgrowth, donde se obtuvieron los siguientes resultados.

Confirma la relación de que el pueblo ladino no estudia en el sector público

Relaciona los departamentos del centro con que no estudian en el sector público, es la primera relación clara de departamento que se verá que en los modelos de predicción toman mucha relevancia.

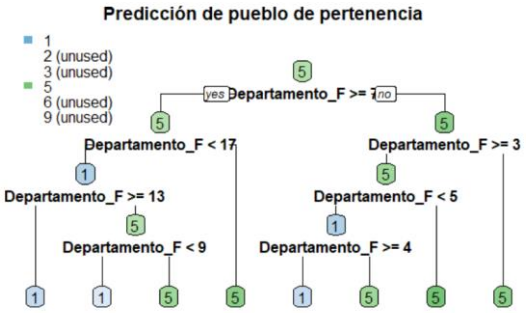
703	{Sector=[2,4]} => {Jornada_Est=[2,5]}	0.4699211	0.6113215	0.9622237	155261

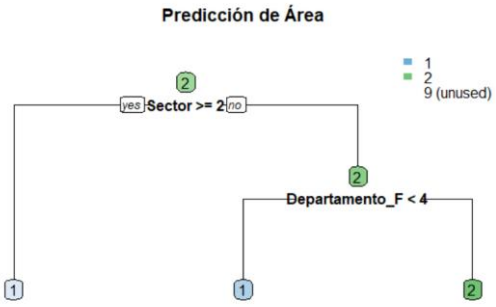
Relaciona que si el sector no es público es probable que la jornada no sea matutina.

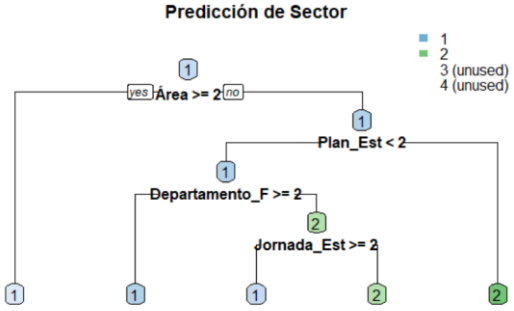
976	{Jornada_Est=[1,2)} => {Pueblo_Per=[5,6]}	0.3100200	0.8501191	1.0708178	102430

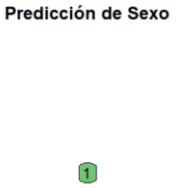
Confirma la relación entre la jornada ser matutina y el pueblo de pertenencia del estudiante es ladino.

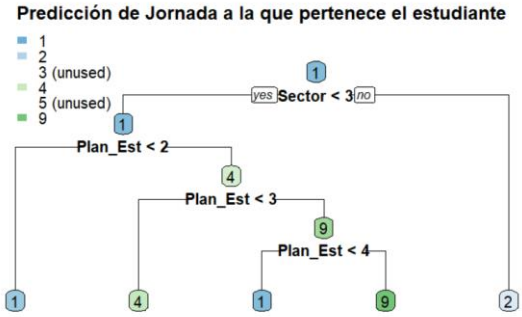
En la segunda fase se realizo los primeros modelos de predicción con los arboles de decisión tomando en cuenta las variables más relacionadas que puedan entregar información importante, entregando los siguientes resultados.











Cómo podemos ver, el departamento es una variable importante para el pueblo de pertenencia, el sector y el área, mientras la jornada depende del sector y el plan de estudio solamente, también en este último nos damos cuentas que efectivamente el sector está ampliamente relacionado con el plan de estudio, donde si el sector no es publico ni privado, la jornada es vespertina, pudiendo significar que los estudiantes de estos sectores estén de algún modo haciendo trabajo infantil y si el sector esta relacionado con el área, el área rural estaría más propenso a caer en enviar a sus hijos a estudiar en jornadas no matutinas.

En el random forest tendremos casi el mismo comportamiento, solo que este si genera un rango más amplio de resultados en algunos casos, por ejemplo, en pueblo de pertenencia sigue existiendo una sobre representación del pueblo maya y ladino. Mientras en sector si esta prediciendo en algunos casos el sector Municipal y Cooperativa a diferencia del árbol de decisión que solo tiene la capacidad de predecir si es público o privado. El análisis de intentar predecir el sexo con la información solo era para ver si había comportamientos extraños o de discriminación en algún sector, área, departamento, pero no fue encontrado ni una relación significativa.

En la última fase se realizó un par de redes neuronales para predecir si la estudiante continua vigente o no según los datos personales, donde se hicieron varios entrenamientos del modelo, obteniendo como resultado alrededor del 87% de acierto.

```
Epoch 1/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8765 - loss: 0.3739 - val_accuracy: 0.8766 - val_loss: 0.3737
Epoch 2/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8763 - loss: 0.3743 - val_accuracy: 0.8766 - val_loss: 0.3736
Epoch 3/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8768 - loss: 0.3732 - val_accuracy: 0.8766 - val_loss: 0.3736
Epoch 4/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8759 - loss: 0.3751 - val_accuracy: 0.8766 - val_loss: 0.3736
Epoch 5/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8764 - loss: 0.3740 - val_accuracy: 0.8766 - val_loss: 0.3737
Epoch 6/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8761 - loss: 0.3745 - val_accuracy: 0.8766 - val_loss: 0.3737
Epoch 7/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 9s 1ms/step - accuracy: 0.8765 - loss: 0.3738 - val_accuracy: 0.8766 - val_loss: 0.3737
Epoch 8/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8767 - loss: 0.3735 - val_accuracy: 0.8766 - val_loss: 0.3736
Epoch 9/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8766 - loss: 0.3737 - val_accuracy: 0.8766 - val_loss: 0.3737
```

Esto se podría decir que es un éxito, y predice de forma adecuada si este estudiante sigue vigente, pero al ingresar diferentes datos en los estudiantes y hacer varias combinaciones la probabilidad de seguir vigente no cambiaba, a lo que se asume que al ser tan amplia la

probabilidad de que siga vigente y despreciable el que no, simplemente tiene un comportamiento similar al predecir con cualquier característica del estudiante.

```
estudiante = np.array([[1, 108, 1, 1, 2, 3, 3, 1, 1, 1, 1, 1]])
p = model.predict(estudiante)
print(p)

estudiante2 = np.array([[16, 1601, 1, 2, 2, 3, 3, 1, 1, 1, 1, 2]])
p = model.predict(estudiante2)
print(p)
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 22ms/step
[[0.8750998]]
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 22ms/step
[[0.8750998]]
```

En la otra red neuronal se volvió a ver sobre el pueblo ladino, si se puede predecir si es ladino o no según sus datos personales, en este si conseguimos un resultado muy interesante. Después de hacer varias iteraciones de entreno del modelo.

```
Epoch 1/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 9s 1ms/step - accuracy: 0.7172 - loss: 0.5279 - val_accuracy: 0.8204 - val_loss: 0.3902
Epoch 2/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8223 - loss: 0.3881 - val_accuracy: 0.8315 - val_loss: 0.3755
Epoch 3/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8299 - loss: 0.3761 - val_accuracy: 0.8375 - val_loss: 0.3696
Epoch 4/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8365 - loss: 0.3697 - val_accuracy: 0.8365 - val_loss: 0.3677
Epoch 5/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8383 - loss: 0.3670 - val_accuracy: 0.8404 - val_loss: 0.3632
Epoch 6/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 8s 1ms/step - accuracy: 0.8400 - loss: 0.3630 - val_accuracy: 0.8380 - val_loss: 0.3599
Epoch 7/20
6769/6769 ━━━━━━━━━━━━━━━━━━━━ 7s 1ms/step - accuracy: 0.8413 - loss: 0.3617 - val_accuracy: 0.8396 - val_loss: 0.3607
Epoch 8/20
```

Se obtuvo alrededor del 85% de precisión, y como la anterior esto no podría significar nada, pero al probar con valores propios este si tiene predicciones de los dos lados, donde valores como el departamento, el sector y el área son muy relevantes para la predicción del estudiante, cómo se había visto en las variables relacionadas en los primero algoritmos.

```
estudiante = np.array([[1, 1, 1, 2, 3, 1]])
p = model.predict(estudiante)
print(p)

estudiante2 = np.array([[2, 2, 2, 1, 1, 16]])
p = model.predict(estudiante2)
print(p)

estudiante3 = np.array([[1, 1, 1, 2, 3, 8]])
p = model.predict(estudiante3)
print(p)
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step
[[0.96537167]]
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step
[[0.3412157]]
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step
[[0.07147411]]
```

Estudiante 1: *Sexo hombre, plan de estudio diario, jornada matutina, área rural, sector municipal, departamento Guatemala, predicción es ladino*
Estudiante 2: Sexo mujer, plan de estudio fin de semana, jornada vespertina, área urbana, sector público, departamento Alta Verapaz, predicción no es ladino
Estudiante 3: Sexo hombre, plan de estudio diario, jornada matutina, área rural, sector municipal, departamento Totonicapán, predicción no es ladino.

Lo más interesante de estas predicciones es que el estudiante uno y tres comparten todas las características menos el departamento, y solo por esta diferencia el modelo cambia la predicción por una amplia diferencia, por lo que se vio en varias pruebas que el centro del país tiene más recursos y una educación bastante diferente a departamentos más alejados del centro.

Conclusiones

Una educación fragmentada en diferentes niveles socioeconómicos es lo que se evidencia con los resultados, se logra ver cómo la educación está demasiado segmentada entre la población y se logra ver las diferentes caras y realidades de Guatemala en la educación que se imparte en el mismo, al final el estudiante termina por ganar los grados y niveles sin importar su nivel socioeconómico, por el mismo marco metodológico de evaluación que esta implementado en el país esto resulta ser complicado el que un estudiante tenga que

recursar el grado, pero aunque los estudiantes no tengan esta complicación si tienen un acceso distinto, donde las áreas rurales tienen mayoritariamente solo acceso al sector público, donde el pueblo ladino tiene mayor acceso a la educación privada, donde se pueden generar asimetrías en grados superiores, donde a nivel universitario se ve los efectos de este desnivel de acceso, donde la mayoría de estudiantes del sector publico son los que no pueden continuar con la educación superior.

En este estudio no se encontró una problemática donde grupos específicos fueran más propensos a reprobar el año académico, pero si se lograron identificar patrones en el acceso a la educación el cuál es un síntomas de estadísticas posteriores en el nivel superior, donde se deben fortalecer la educación publica en estos niveles para que sea una opción interesante y buena para todo guatemalteco y no solo para los grupos subalternos del país que terminan por conformarse por la educación a veces pobre que se imparte.

Referencias

INE (2023). Datos y estadísticas de la educación superior en Guatemala. INE https://www.ine.gob.gt/educacion/

INE. República de Guatemala: Compendio Estadístico de Educación 2013. https://www.ine.gob.gt/sistema/uploads/2015/09/17/35aftsjzEBB6YMPlOcRdUF3SVqTmbAnW.pdf

Instituto Internacional de la UNESCO para la Educación Superior en América Latina y el Caribe (2023). El derecho a la educación superior en América Latina y el Caribe: compendio de notas conceptuales. UNESCO ED/HE/IESALC/IP/2023/25. https://unesdoc.unesco.org/ark:/48223/pf0000387656_spa.locale=es

González Orellana, C. (2007). Historia de la educación en Guatemala. Guatemala: Editorial Universitaria, Universidad de San Carlos de Guatemala.

Lemus Cedillo, Carla Beatriz (2017). La educación [PDF]. Recuperado 03 de abril de 2020 en https://gradoceroprensa./2017/01/02/la-educación/

Menéndez, L. A. (2006). La educación en Guatemala, 1954-2004: enfoque histórico-estadístico. Guatemala: Editorial Universitaria, Universidad de San Carlos de Guatemala.

Ministerio de Educación. Acuerdo ministerial: sobre la evaluación. https://edu.mineduc.gob.gt/DIGEESP/documents/Acuerdo%20Ministerial%201171-2010%20Reglamento%20de%20Evaluaci%C3%B3n.pdf

UNESCO (2019). Los futuros de la educación: aprender a convertirse. UNESCO, documento de programa. https://unesdoc.unesco.org/ark:/48223/pf0000370801_spa

UNESCO para la Educación Superior en América Latina y el Caribe (1990) Declaración Mundial sobre Educación para Todos y Marco de Acción para Satisfacer las Necesidades Básicas de Aprendizaje. https://unesdoc.unesco.org/ark:/48223/pf0000127583_spa

SITEAL, UNESCO. Resumen del marco normativo y estructura del sistema educativo nacional, Guatemala. https://siteal.iiep.unesco.org/pais/educacion-pdf/guatemala