

Modelo de aprendizaje automático para el análisis predictivo de las habilidades básicas del pensamiento

Jerónimo Aparicio Juárez

Instituto Tecnológico Superior de Teziutlán

Maestría en Sistemas Computacionales

Aire Libre, Fracción I y II, 73960, Teziutlán, Puebla.

jeronimoaj@gmail.com

Resumen

En la presente investigación se abordará un subcampo de las ciencias de la computación y a su vez una rama de la inteligencia artificial denominado aprendizaje automático, además se aplicarán técnicas de minería de datos con la finalidad de tomar decisiones en base a predicciones de los factores que inciden en el dominio de las habilidades básicas del pensamiento, aplicando árboles de clasificación. Para cumplir el objetivo se analizó desde diferentes fuentes de información distintas colecciones de datos con la finalidad de parametrizar un modelo que permita resolver la problemática de las causas que inciden en el bajo rendimiento escolar. Es importante mencionar que en la mayoría de los casos fue necesario realizar una estadística previa de la información recabada para poder formar los elementos que incluyen nuestra línea de investigación.

Palabras clave: Aprendizaje, Automático, arboles, decisión, habilidades, básicas, pensamiento

Abstract

This research will address a subfield of computer science and in turn a branch of artificial intelligence, machine learning, in addition data mining techniques will be applied with the proposal to make decisions based on predictions of the factors that they influence the mastery of basic thinking skills, applying classification trees. To meet the objective, different data collections were analyzed from different sources of information with the specification of parameterizing a model that allows solving the problem of the causes that affect low school performance. It is important to include that in the majority of cases it was necessary to perform a prior statistic of the information collected in order to form the elements that include our line of research.

Keywords: Learning, Automatic, trees, decision, skills, basic, thinking

2. Introducción

En la actualidad las tendencias tecnológicas propician nuevas formas de darle tratamiento a los datos, es por esta razón que hoy en día no es suficiente solo almacenar colecciones de datos para su análisis rudimentario, sino que se requiere realizar predicciones que permitan alertar de manera oportuna y precisa situaciones críticas, así mismo, es necesario descubrir hechos contenidos en las bases de datos sin ninguna intervención del humano, pero con el objetivo de servir de apoyo a la toma de decisiones.

En el contexto descrito anteriormente, el presente proyecto se sustenta y propone dar solución a una problemática presente en el dominio de las habilidades básicas del pensamiento en educación básica las cuales inciden en índices de reprobación, bajo rendimiento y deserción escolar por situaciones no conocibles, pero si observables en los resultados obtenidos, los cuales solo pueden predecirse con la intervención de un humano, pero de forma muy carente y no eficaz. Por esta razón, se creará un aplicativo automatizado que muestre predicciones antes de que los eventos sucedan.

La línea de investigación que se abordará para tratar la problemática planteada es denominada Aprendizaje Automático, con ella se desarrollará nuestro modelo de predicción, así como también, se aplicaran técnicas de minería de datos con el objetivo de canalizar métodos estadísticos rudimentarios en los procesos que actualmente presentan problemas de eficiencia y escalabilidad.

"Las 'leyes del pensamiento' no solo dependen de las propiedades de las células cerebrales, sino del modo en que están conectadas"

-Marvin Lee Minsky, científico estadounidense.

¹Marvin Lee Minsky, es considerado uno de los padres de la inteligencia artificial. Fue cofundador del laboratorio de inteligencia artificial del Instituto Tecnológico de Massachusetts o MIT.

1. Planteamiento del problema

Las escuelas de educación básica presentan un problema para la toma de decisiones de forma oportuna en lo referente a las habilidades básicas del pensamiento en los alumnos, las cuales están comprendidas en habilidades de escritura, lectura y matemáticas.

Para coleccionar datos sobre esas problemáticas la Secretaría de Educación Pública pone a su alcance el Sistema de Alerta Temprana (SiSAT) en el cual el cuerpo docente, administrativo y directivo suministra información sensible que posteriormente es proyectada en indicadores y gráficas en el sistema, sin embargo, este no realiza ninguna predicción oportuna sobre los datos conocibles, sino que, por el contrario, en algunos casos presenta inconsistencias en:

- La graficación visual de los datos
- Las escuelas presentan inconformidad en la veracidad del valor numérico que se asigna al alumno en determinado componente de lectura y escritura.
- Debido a lo anterior, las problemáticas tratadas en los Consejos Técnicos Escolares recaen en los temas de:
- Índices de reprobación
- Cumplir con la normalidad mínima
- Mejorar el resultado de las evaluaciones de SiSAT el cual no permite predecir los siguientes rubros:
- Identificar los factores que inciden en el avance y retroceso del alumno en la lectura.
- Identificar los factores que dificultan la adquisición del cálculo mental.
- Identificar los factores que interfieren en la redacción de textos.

Por lo tanto, se requiere el desarrollo de un modelo que pueda ser aplicable a un sistema y que permita el manejo de la información de forma eficiente para lograr un resultado eficaz que dé solución a lo descrito anteriormente

3. Justificación

El modelo de Aprendizaje Automático que se desarrollará a continuación permitirá dar predicciones con respecto a los procesos de adquisición de la lectura, escritura y cálculo mental que en la actualidad no existen como herramienta de apoyo en las instituciones de educación básica, al implementar este modelo los colegios podrán ejecutar un plan de contingencia basado en resultados significativos, pertinentes, factibles y oportunos.

El impacto social del modelo propiciará un beneficio colectivo para los alumnos y el centro escolar que se verá reflejado en el aumento de la calidad educativa del plantel a través del cambio en las metodologías de la práctica docente actual por la implementación de estrategias e instrumentación didáctica innovadoras que permitan al alumno desarrollarse en un entorno de aprendizaje favorable y armónico. Así mismo, se eliminará el uso exagerado de documentos en papel debido a que la aplicación del modelo en un sistema integral ofrecerá movilidad vía web y móvil para su implementación en producción.

El sistema permitirá automatizar procesos predictivos de forma rápida y confiable apoyando a la solución de problemas prácticos presentes en el centro escolar, haciendo uso de algoritmos de inteligencia artificial que han marcado las últimas tendencias tecnológicas. Así como también, será adaptable al nuevo modelo educativo.

Al mitigar las problemáticas que influyen en el proceso de adquisición de las habilidades básicas del pensamiento, el sistema aportará beneficios a la Ruta de Mejora Escolar (RME), no solo de manera local, sino nacional en caso de ser aceptada como propuesta factible en los centros educativos del sistema básico a través de la Secretaría de Educación Pública.

El aprendizaje automático y minería de datos aplicados en la implementación del modelo permitirá generalizar comportamientos a través de ejemplos que con los métodos rudimentarios no se podría dar solución de forma inmediata; permitiendo realizar un estudio preciso de los alumnos con bajo rendimiento escolar que se encuentra en estado crítico.

Por otra parte, la minería de datos, relacionará variables que extraerán información valiosa y necesaria, sin la

la intervención de un humano, pero para beneficio de este. Dando como resultado predicciones anticipadas y confiables.

Objetivos generales

Desarrollar un modelo de predicción de las Habilidades Básicas del Pensamiento, para conocer sus factores de incidencia en los colegios de nivel básico aplicando aprendizaje automático y técnicas de minería de datos, que permita tomar decisiones para crear una Ruta de Mejora Escolar dinámica para que anticipadamente los colegios puedan realizar una intervención pedagógica para tratar los resultados de predicción y evitar que un grupo llegue a un estado crítico que sea más difícil de tratar y genere más costos.

Específicos

- Predecir las habilidades básicas del pensamiento:
 - Lectura
 - Cálculo mental
 - Redacción de textos
- Representar de forma gráfica y predictiva los factores que inciden en:
 - Avance y retroceso en la lectura
 - Dificultad en el cálculo mental
 - Redacción de textos

Hipótesis

Variable independiente

A mayor número de ejemplos en el modelo de aprendizaje automático

Variable dependiente

Mayor exactitud y se evitarán problemas de overfitting u underfitting en las predicciones del sistema para mejorar las habilidades básicas del pensamiento de forma anticipada.

Hipótesis resultante

“A mayor número de ejemplos en el modelo de aprendizaje automático mayor exactitud y se evitarán problemas de overfitting u underfitting en las predicciones del sistema para mejorar las habilidades básicas del pensamiento de forma anticipada”.

4. Metodología

Tecnologías de programación y sistemas

El escenario de prueba que se utilizó es el siguiente:

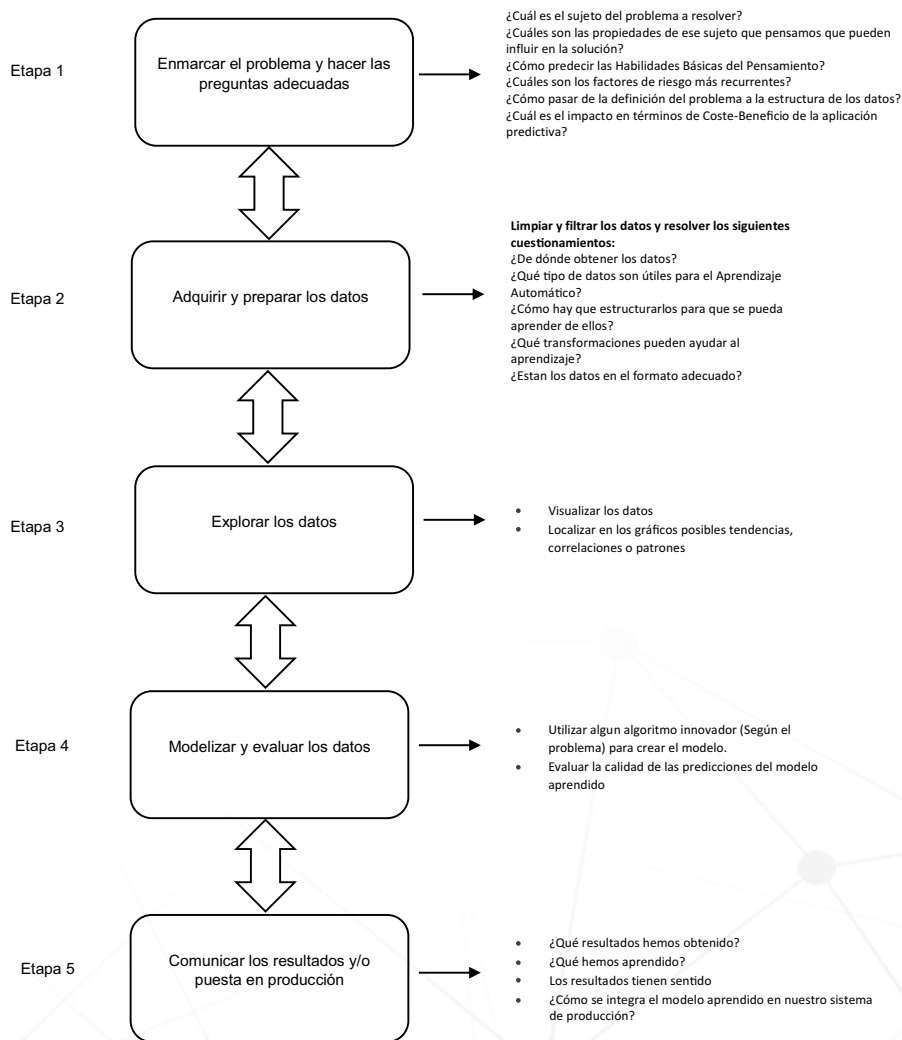
- Sistema operativo Windows 10 Versión 1909 (10.0.18363.476)
- Equipo portátil Acer Intel(R) Core(TM) i7-4810MQ CPU 2.80 Ghz, RAM 8GB, SSD 240 GB

Para realizar el modelo de Aprendizaje Automático (AA) se utilizó por su versatilidad y potencia el lenguaje de programación Python 3.8.0, junto con librerías adicionales que se presentan a continuación:

Nombre	Funciones
Pandas 0.25.3	• Leer datasets, • Realizar estadísticas básicas, • Concatenar datasets, etc.
NumPy 1.16.2	• Cálculo numérico, • álgebra lineal, etc.
Matplotlib 3.0.3	• Permite realizar representaciones gráficas • Histogramas, • Gráfico de barras, • Mapas de calor, etc.
Scikit-learn 0.21.1	Biblioteca de AA ampliamente eficiente en: • Ciencia de datos • Combina la carga de datos, representación, cálculo matricial • Tiene implementado métodos clásicos (regresión lineal, árbol de decisión, clustering, etc.) • Nos da formas de predecir el valor de salida basado en el modelo de partida y medir la eficiencia.

Para la instalación de las librerías se utilizó el instalador de paquetes PIP en su versión 19.3.1.

Etapas en el análisis de datos



Las flechas con doble sentido indican que en cualquier momento en el análisis de datos puede ser posible no solo ir hacia delante, sino regresar una etapa atrás y replantearnos algo que no hayamos tenido en consideración.

Población y muestra

La investigación se realizó en un colegio perteneciente al estado de Veracruz, a alumnos y docentes de educación primaria estatal pertenecientes a una zona escolar que la integran 48 escuelas, de la cual se tomó una muestra de 329 alumnos y 15 docentes dando un total de la muestra de 344 personas examinadas.

Grado/Grupo	Número de alumnos examinados
1a	28
1b	29
2a	27
2b	24
3a	21
3b	23
3c	22
4a	26
4b	26
5a	30
5b	29
6a	23
6b	21

Total de alumnos	329
Total docentes	15

Total de la muestra	344
---------------------	-----

Para la realización de esta demostración solo se utilizará una muestra parcial correspondiente al grupo de alumnos de primero "A". Y se analizarán los factores de incidencia en las Habilidades Básicas del Pensamiento.

Desarrollo

Para poder tener un contexto amplio de nuestro análisis es importante extraer los siguientes datos en Python:

```

Información del dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28 entries, 0 to 27
Data columns (total 14 columns):
ID                28 non-null int64
Nombre            28 non-null object
Estilos_Aprendizaje 28 non-null object
Madurez_Intelectual 28 non-null object
Relaciones        28 non-null object
Situacion_Emocional 28 non-null object
Estado_Salud      28 non-null object
Situacion_Economica 28 non-null object
Situacion_Familiar 28 non-null object
Asistencia        28 non-null object
Participacion     28 non-null object
Lectura           28 non-null object
Escritura          28 non-null object
Calculo_Mental    28 non-null object
dtypes: int64(1), object(13)
memory usage: 3.1+ KB
None

```

Como podemos observar para nuestro análisis contamos con 28 instancias, esta prueba se realizó con valores categóricos tanto para los datos como para las etiquetas.

El valor ID se descarta, para nuestro análisis de datos, pues no tiene ningún valor predictor.

La cantidad de memoria de uso de nuestra fuente de información es: 3.1+ Kilobytes.

Es importante mencionar que para nuestro análisis de datos omitiremos los datos de las columnas correspondiente al ID y Nombre, ya que no tiene una propiedad predictora. Dando como resultado 12 características observadas en 28 alumnos examinados.

Exploración de datos previos al análisis

Describir nuestro dataset es muy importante para conocer algunos indicadores importantes como puede ser conocer la cantidad de elementos por columna examinados, media, desviación estándar, valores máximo y mínimo, entre otros.

	Estilos_Aprendizaje	Madurez_Intelectual	Relaciones	Situacion_Emocional	Estado_Salud	Situacion_Economica	Situacion_Familiar	Asistencia	Participacion
count	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.0	28.000
mean	2.250000	2.714286	1.035714	1.142857	0.854286	1.500000	0.821429	1.0	0.857
std	0.927961	0.712697	0.331343	0.440395	0.189982	0.57735	0.390021	0.0	0.705
min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.0	0.000
25%	1.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.0	0.000
50%	3.000000	3.000000	1.000000	1.000000	1.000000	2.000000	1.000000	1.0	1.000
75%	3.000000	3.000000	1.000000	1.000000	1.000000	2.000000	1.000000	1.0	1.000
max	3.000000	4.000000	2.000000	2.000000	1.000000	2.000000	1.000000	1.0	2.000

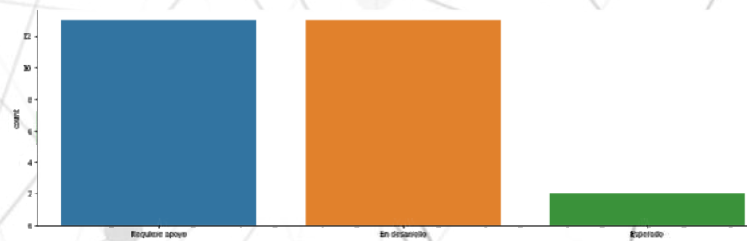
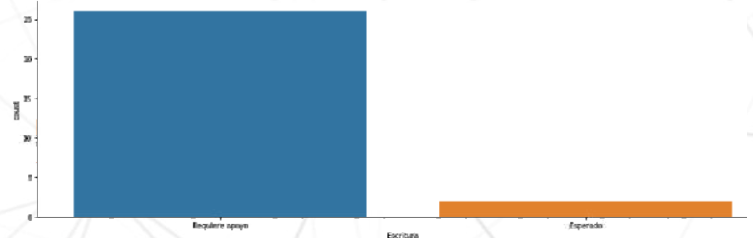
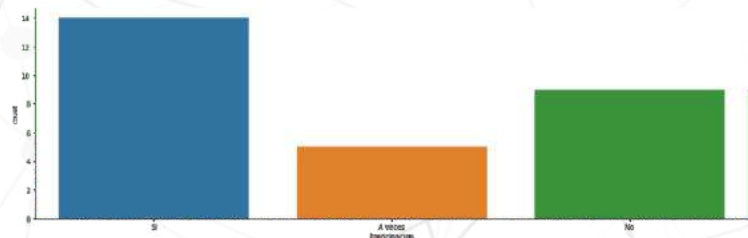
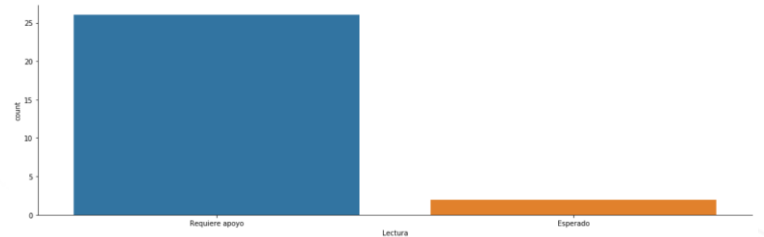
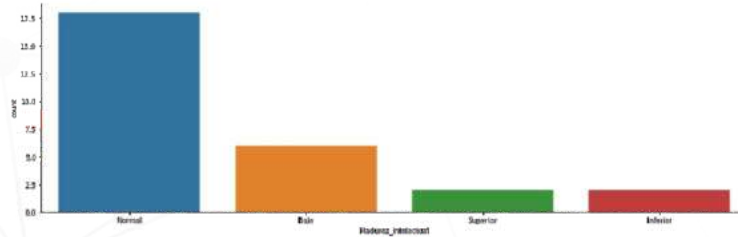
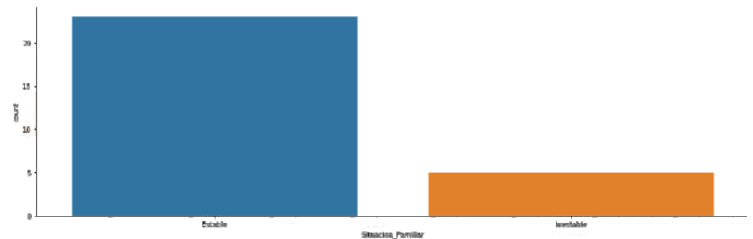
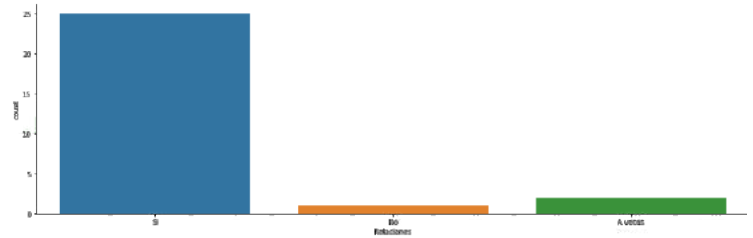
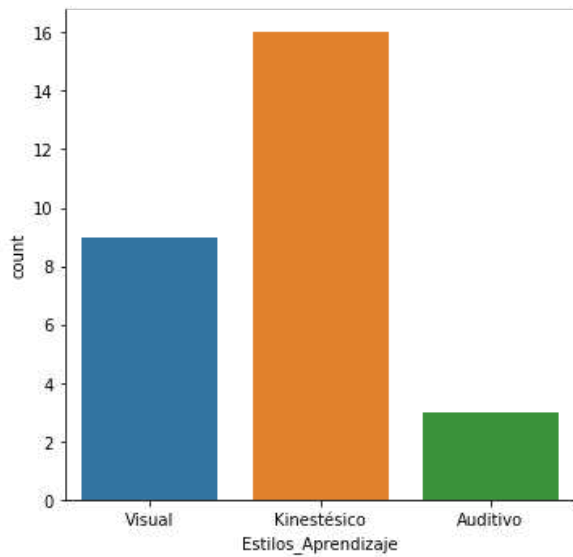
```

DISTRIBUCIÓN DE HABILIDADES BÁSICAS DEL PENSAMIENTO (HBP) | GRUPO: 1B
HBP de LECTURA
Lectura
Esperado                2
Requiere apoyo         26
dtype: int64
HBP de ESCRITURA
Escritura
Esperado                2
Requiere apoyo         26
dtype: int64
HBP de CÁLCULO MENTAL
Calculo_Mental
En desarrollo           13
Esperado                2
Requiere apoyo         13
dtype: int64

```

En estos resultados observamos la distribución de las HBP, podemos observar que para las primeras 2 correspondientes a "Lectura" y "Escritura", no se tiene a ningún alumno en el nivel de "En desarrollo", lo contrario a la HBP de "Cálculo Mental" que integra una distribución de alumnos en sus 3 respectivos niveles.

Exploración gráfica de la estadística previa de los datos



Modelo de aprendizaje automático supervisado: Árboles de decisión

Los árboles de decisión se aplican a dos contextos diferentes:

La regresión para la predicción sobre valores continuos y la clasificación que es la que se utilizará para nuestro modelo de predicción debido a las características de la problemática planteada al inicio del presente.

El árbol de clasificación que se configuro en Python toma en cuenta las siguientes premisas:

① Se utilizó un archivo de datos con información de los resultados obtenidos en el examen de diagnóstico sobre las HBP en alumnos de Primero B de Educación Primaria.

② Las variables dependientes de las HBP (Lectura, Escritura y Cálculo mental) son: Esperado, En desarrollo y Requiere apoyo. Los datos de las variables independientes son: Estilos de aprendizaje, Madurez intelectual, Relaciones, Situación emocional, Situación familiar, Asistencia, Participación.

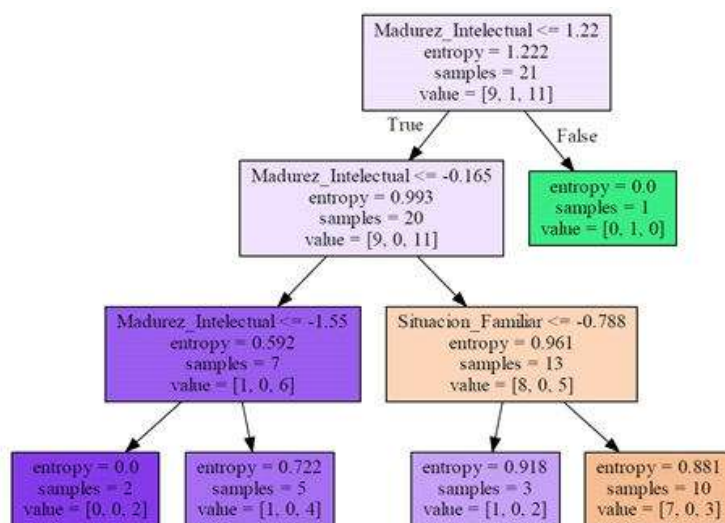
③ Utilizaremos la Madurez Intelectual y la Situación Familiar para crear nuestro árbol de clasificación con las etiquetas Esperado, En Desarrollo y Requiere apoyo y con ello poder clasificar registros nuevos. Para ello crearemos el árbol con el conjunto de datos a los que les llamaremos conjunto de entrenamiento.

④ Nuestro árbol de clasificación se ejecutó con 28 registros de datos históricos de alumnos de los cuales 25% se utilizaron para probar el modelo y 75% para entrenarlo o crear el árbol bajo los criterios de Madurez Intelectual y Situación Familiar.

A continuación, se presenta el árbol de clasificación generado del cual se recomienda analizar lo siguiente:

- **Condición:** Es un nodo donde se toma alguna decisión
- **Entropía:** Es una medida de impureza.
- **Samples:** Número de muestras que satisfacen las condiciones necesarias para llegar a este nodo

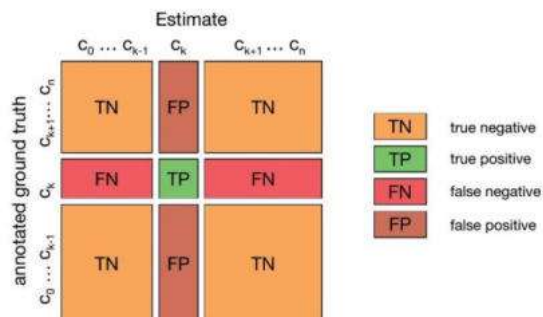
- **Value:** Cuántas muestras de cada clase llegan a este nodo



Con una matriz de confusión

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Que representa lo siguiente:



Como podemos darnos cuenta tenemos predicción verdadero positivo, 6 verdadero negativo

5. Resultados y discusión

Considerando las características únicas de nuestros datos se seleccionó evaluar con nuestro modelo de aprendizaje automático la Habilidad Básica del Pensamiento de Cálculo Mental que cuenta con 3 etiquetas respectivas.

La interpretación de nuestro clasificador se realizó de la siguiente forma:

En el nodo final ubicado del lado derecho en el segundo nivel, dice entropía = 0, muestras = 1 y value = [0, 1, 0] es un nodo de color verde que indica que los alumnos de Madurez Intelectual mayor a -0.165 o

Es importante aplicar estrategias para resolver desequilibrio de datos en Python con la librería imbalanced-learn debido a pues como nos dimos cuenta al momento de explorar nuestros datos previamente, en nuestro conjunto de datos de entrenamiento contamos con que algunas de las clases de muestra es una clase “minoritaria”, lo cual quiere decir que tenemos poquitas muestras o ejemplos de entrenamiento. Esto provoca un desbalanceo en los datos que se utilizan para el entrenamiento de nuestro modelo de aprendizaje automático.

Finalmente es importante mencionar que a mayor número de ejemplos nuestro modelo tendrá una mayor precisión en las predicciones, evitando en menor medida tener problemas de overfitting u underfitting que pueden afectar nuestro modelo al no generalizar correctamente los datos de entrenamiento o cualquier dato del dominio del problema.

Se debe tener en cuenta que los árboles de decisión tienen tendencia al sobreajuste.

6. Trabajo Futuro

El modelo de aprendizaje automático se implementará en un sistema integral donde se puedan ingresar los diagnósticos generales por grupo y grado, así como también los diagnósticos complementarios para realizar las predicciones.

Se implementará los siguientes casos de uso para realizar las visualizaciones de:

- Predicción de las Habilidades Básicas del Pensamiento (HBP) por grupo y grado.
- Factores que inciden en las Habilidades Básicas del Pensamiento
- Resultados por grupos paralelos

Para satisfacer los requerimientos de los actores principales de la institución educativa Docente, Directivo y Supervisor Escolar.

Es importante mencionar que existen otros factores que inciden las HBP de los alumnos, como lo es la “Metodología del docente”. Por esta razón, en trabajos posteriores se realizará una concatenación de datasets que ya incluyan estos factores de las muestras recolectadas.

Otro aspecto importante que se realizará es la aplicación de diferentes tipos algoritmos clasificadores que permita comparar los resultados y evaluarlos, para decidir cuál es el más óptimo.

7. Conclusiones

Como pudimos observar a través de la creación del modelo, el análisis de datos predictivos resulta ser de gran utilidad en la toma de decisiones en los planteles educativos. Por esta razón es de vital importancia que exista un financiamiento a la investigación en el área de la educación para consolidar esfuerzos en la prevención del bajo rendimiento escolar aplicando algoritmos de inteligencia artificial en todas sus ramas que permitan impulsar el mejoramiento de la calidad educativa.

8. Referencias bibliográficas

Günther, C. W. (2010). Process mining: A two-step approach to balance between underfitting and overfitting. En Software and Systems Modeling (Vol. 9). <https://doi.org/10.1007/s10270-008-0106-z>