**INSTITUTO POLITECNICNO NACIONAL**

**ESCUELA SUPERIOR DE COMPUTO**

**MINERIA DE DATOS**

**ESCOM**

PROYECTO ARBOLES

“**CLASIFICACION Y LIMPIEZA DE DATOS CON C4.5”**

*PRESENTA*

**MARTINEZ LOPEZ SEBASTIAN**

[smartinezl1301@alumno.ipn.mx](mailto:smartinezl1301@alumno.ipn.mx) | 6CV1

**RAMIREZ RESENDIZ LUIS ROQUE**

[lramirezr1602@alumno.ipn.mx](mailto:lramirezr1602@alumno.ipn.mx) | 6CV1

**RODRIGUEZ YEPEZ JAAZIEL**

[jrodriguezy1600@alumno.ipn.mx](mailto:jrodriguezy1600@alumno.ipn.mx) | 6CV1

*PROFESOR*

**Dr. FABIOLA OCAMPO BOTELLO**

ÍNDICE

[RESUMEN 3](#_Toc122655032)

[PALABRAS CLAVE 3](#_Toc122655033)

[INTRODUCCION 4](#_Toc122655034)

[OBJETIVO 4](#_Toc122655035)

[DESCRIPCION DEL ESTUDIO 4](#_Toc122655036)

[DICCIONARIO DE DATOS 5](#_Toc122655037)

[DESARROLLO DEL PROYECTO. 6](#_Toc122655038)

[Carga de los datos. 6](#_Toc122655039)

[Preparación del analisis 7](#_Toc122655040)

[Arbol de clasificación: 8](#_Toc122655041)

[Matriz de confusión: 10](#_Toc122655042)

# RESUMEN

Uno de los principales retos que enfrentan los científicos de datos es la elección del arbol de decisión a emplear en su proceso KDD, debido a que existen múltiples arboles que pueden satisfacer un problema dado, sin embargo, es tarea del científico identificar e implementar la mas adecuada para resolver los objetivos propuestos con el mayor margen de ganancia de informacion. Dentro de la practica propuesta, elegimos un dataset, acerca del aprovechamiento escolar que tienen alumnos durante su periodo de exámenes, los cuales se verán afectados por las variables como género, grado máximo de estudios de los padres, el tipo de alimentación, si presentaron algún curso y las calificaciones en sus asignaturas de matemáticas, lectura y escritura, para analizar la informacion se utilizara el arbol ID3, el cual nos permitirá clasificar bajo que condiciones un alumno es sobresaliente y donde no.

# PALABRAS CLAVE

Arbol ID3, Aprovechamiento escolar, Clasificación de datos, Minería de datos.

# INTRODUCCION

Uno de los principales problemas que puede afectar el aprovechamiento del alumno, es en gran medida las herramientas que este posee para la adquisición de nuevo conocimiento, no es una mentira que el desempeño académico de los estudiantes se ve afectado enormemente si estos carecen de herramientas que pueden ayudarles a aprovechar al máximo su capacidad, dentro de este rubro, en muchas ocasiones afecta de gran manera la alimentación de los mismos, o el grado de estudios que tienen sus progenitores, asi mismo, aunque no se ve ligado en la práctica, es nivel socioeconómico familiar, es un detonante para la adquisición de nuevas herramientas.

Sin embargo, el estudio de este dataset nos puede permitir observar patrones demográficos en el comportamiento de las personas y a su vez, brindarle a un especialista la informacion necesaria para que este mismo pueda crear nuevas estrategias de aprendizaje para los grupos más vulnerables.

# OBJETIVO

* Identificar y clasificar las calificaciones promedio que deben tener los alumnos en las asignaturas de Matemáticas, Lectura y Gramática para ser catalogados por “Buenos estudiantes”, “Estudiantes promedio” y “Malos estudiantes”.

# DESCRIPCION DEL ESTUDIO

Generar un arbol de clasificación, el cual, pueda delimitar que calificaciones son las ideales para catalogar a un estudiante como “Bueno”, “Promedio” o “Malo”. Además, se busca observar, experimentar y analizar como funciona el arbol C4.5, ¿Cómo se implementa? Y ¿Cómo se comporta con el dataset propuesto?

# DICCIONARIO DE DATOS

Dentro del dataset, podemos observar que se encuentran las siguientes variables, asi mismo, en el siguiente apartado, las desglosamos para la comprensión del lector.

**Tabla 1: Diccionario de los datos**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Descripción** | **Dominio de valores** | **Tipo de dato** |
| Género | Género del alumno | Masculino  Femenino | Nominal |
| Grupo Étnico | Grupo al cual pertenece | Grupo A  Grupo B  Grupo C  Grupo D  Grupo E | Nominal |
| Educación  Máxima padres | Grado máximo de estudios de los padres | Secundaria Secundaria incompleta Universidad incompleta Titulo técnico Licenciatura Maestría | Ordinal |
| Desayuno | El desayuno que se les da a los alumnos | Estándar  Gratis-Reducido | Nominal |
| Curso de preparación | Si estuvo inscrito en un curso de reforzamiento de conocimientos | Completado  No completado | Nominal |
| Calificación  matemáticas | Puntaje obtenido en la asignatura de matemáticas | 0-100 | Cuantitativo |
| Calificación  lectura | Puntaje obtenido en la asignatura de lectura | 0-100 | Cuantitativo |
| Calificación gramatica | Puntaje obtenido en la asignatura de escritura | 0-100 | Cuantitativo |

# DESARROLLO DEL PROYECTO.

## Carga de los datos.

Al iniciar nuestra carga de datos, nos percatamos que el lenguaje de los mismos estaba en Ingles, como se muestra en la Ilustración 1, por lo que, se decidió hacer una limpieza de los datos, transformándolos a Español, asi mismo, limpiando los posibles datos nulos, o con poca informacion y que estos no afecten el analisis de los mismos.

Tabla

Descripción generada automáticamente

**Ilustración 1: Carga de datos**

Una vez completado el proceso de limpieza, los datos quedaron estructurados de la siguiente manera:

Imagen que contiene Calendario

Descripción generada automáticamente

**Ilustración 2: Datos transformados y limpios**

## Preparación del analisis

Para poder categorizar a los alumnos es necesario calcular los promedios de estos, es por eso que se suman las calificaciones de las 3 materias de los alumnos, se dividen entre 3 y por último, se les asigna una etiqueta acorde a su promedio, esto para poder categorizarlos en “Bueno”, “Promedio” y “Malo”. Asi mismo, estas son las reglas para categorizarlos

Diagrama

Descripción generada automáticamente con confianza media

**Ilustración 3: Reglas para el promedio**

Gracias a esto, nuestra data queda estructurada de la siguiente manera:

Interfaz de usuario gráfica, Tabla, Excel

Descripción generada automáticamente

**Ilustración 4: Data lista**

Una vez con la data asi, procedemos a crear el arbol, para identificar que reglas está siguiendo para saber el tipo de estudiante.

## Arbol de clasificación:

Una vez creado el arbol, nos arrojó el siguiente diagrama, el cual, por motivos de longitud, tuvimos que recortar un poco para que se pueda identificar en el reporte.

Diagrama

Descripción generada automáticamente

**Ilustración 5: Arbol generado**

En la Ilustración 5, podemos observar que se tiene un 28.7% de alumnos “Malos”, un 65.4% de alumnos regulares y un 5.9% de alumnos excelentes, los cuales nos dan un total de 700 alumnos para la prueba del arbol, y que, a su vez, la ira clasificando acorde a sus calificaciones, empezado por la calificación de gramática, donde es la que le genera mas ganancia de informacion para poder hacer la partición de nodo a nodo.

Podemos observar también que si la calificación de gramática esta por debajo del 58.8, los estudiantes son mas propensos a ser catalogados por “Malos” y que solamente un grupo de 14 estudiantes pueden salvarse de esa categoría si su calificación en matemáticas es superior a 64.5 o bien que su calificación en gramática este entre los rangos de 56.5 a 64.5.

Por otra parte, en el arbol también se puede llegar a apreciar que en dado caso de que dos en las materias de gramática y matemáticas tengan un puntaje superior a 9, estos son catalogados automáticamente como estudiantes “Buenos”.

Por último, si los estudiantes tiene un rango de calificación del 60 al 88, son catalogados como estudiantes “Promedio”. A continuación, se presentan las reglas que creo el modelo.

Tabla

Descripción generada automáticamente con confianza media

**Ilustración 6:Reglas de clasificación**

## Matriz de confusión:

Asi mismo, la matriz de confusión obtenida es la siguiente:

Interfaz de usuario gráfica, Texto, Aplicación, Tabla

Descripción generada automáticamente

**Ilustración 7:Matriz de confusión**

Donde podemos observar que nuestra tasa de éxito es del 95% y nuestra tasa de error es de 5%, a su vez, tenemos un total de 285 alumnos clasificados correctamente y 15 clasificados incorrectamente.