Predicción de éxito estudiantil con árboles de decisión.

Sara Rodríguez Velásquez
Stiven Yepes Vanegas
Medellín, 02/06/2020

Estructuras de Datos Diseñada

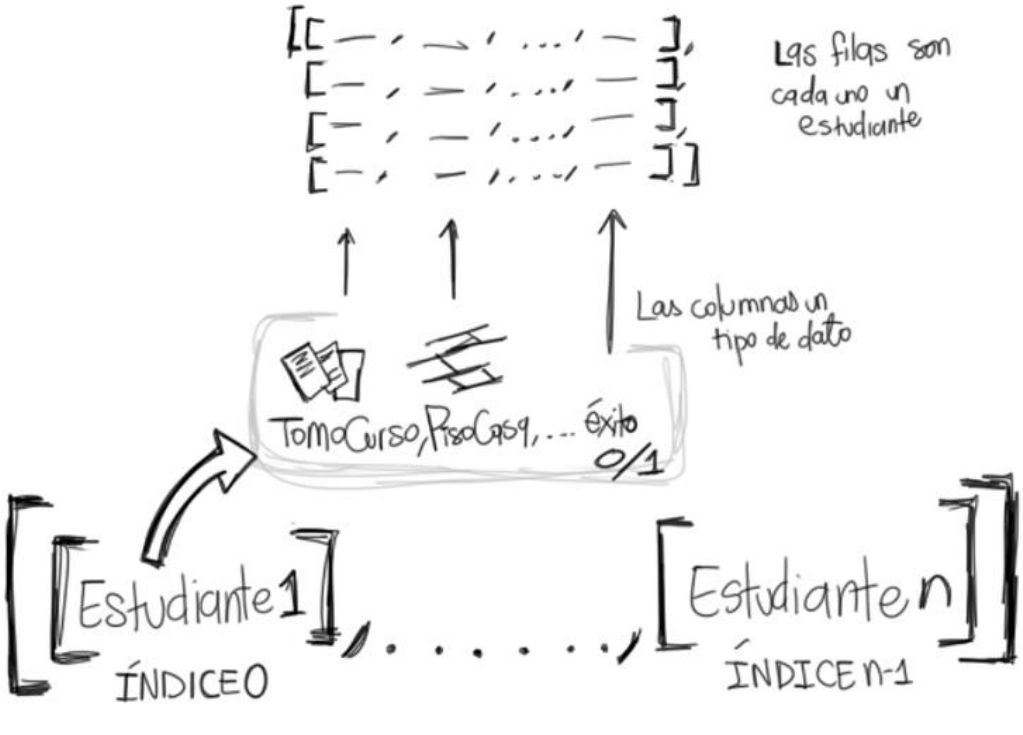


Gráfico 1: Arreglo numpy con datos de personas agrupados en cada element (element) de este.

Operaciones de la Estructura de Datos

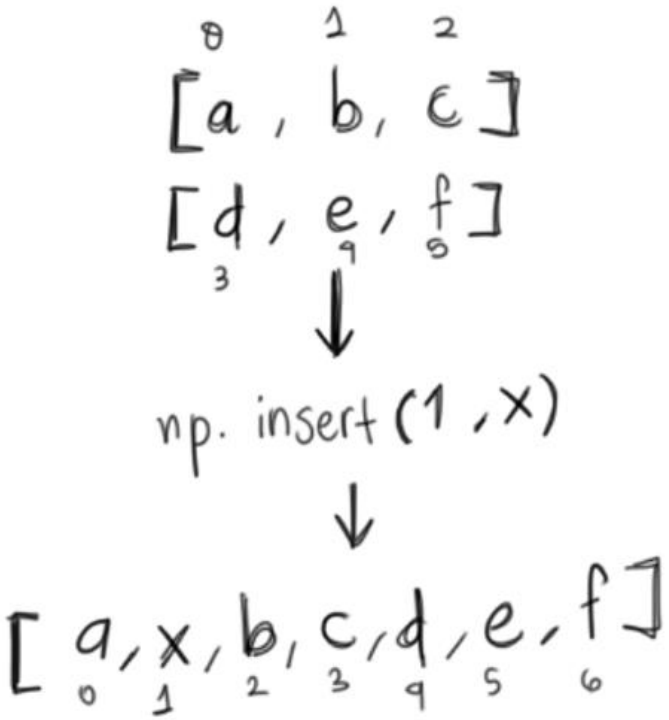
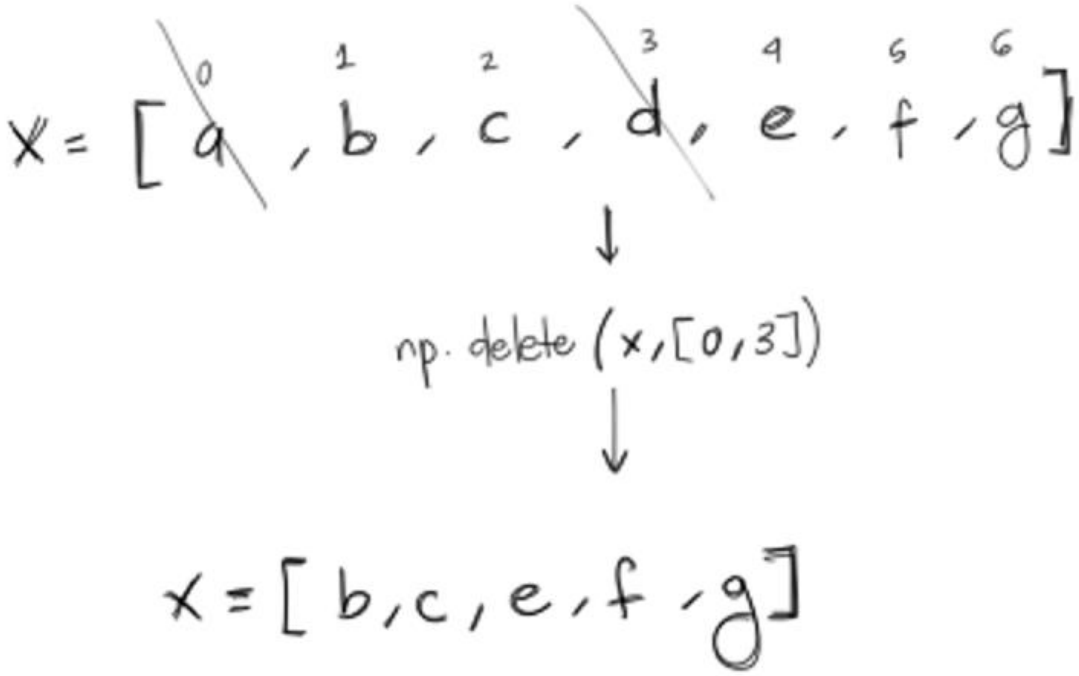


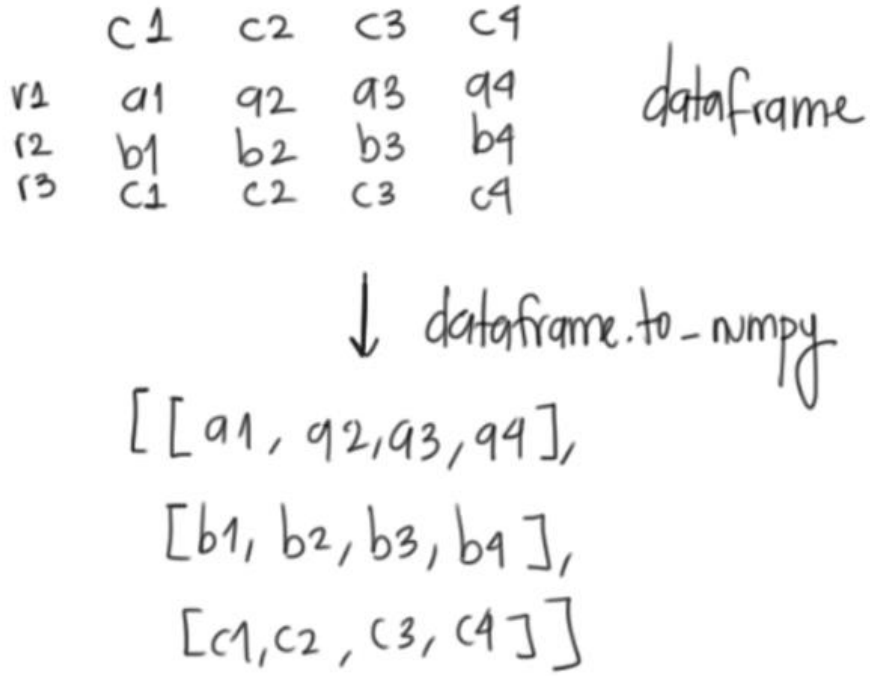
Gráfico 2: Inserción de un dato x en un arreglo de letras indicando solo la posición.

Operaciones de la Estructura de Datos



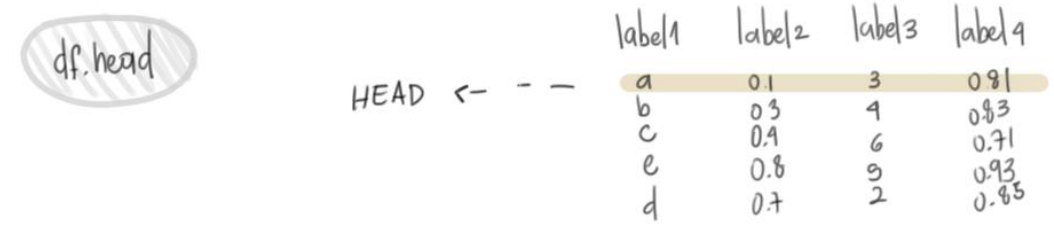
Gráfica 3: Borrado de 2 datos de un arreglo de letras, indicando la posición de borrado.

Operaciones de la Estructura de Datos



Gráfica 4: Conversión de un formato de pandas dataset a un arreglo numpy.

Operaciones de la Estructura de Datos



dataframe.head = Operación para retornar los primeros elementos de un dataframe.

dataframe.tail = Operación para retornar los últimos elementos de un dataframe.

Gráfica 5: Separación de cabeza y cola de un dataframe de Pandas.

Operaciones de la Estructura de Datos

dataframe.columns = retorna una lista de python de los títulos de el dataframe y su tipo.

Gráfica 6: Separación de los índices, o labels del dataframe.

Complejidad de las operaciones de la estructutra de datos

Método	Complejidad
Función view (Ver elemento)	O(1)
Crear array vacío (numpy.empty)	O(1)
Crear array con ceros (numpy.zeros)	O(n)
Crear array multidimensional con ceros (numpy.zeros)	O(nm), n= filas, m=columnas
Función Copy (Copiar array)	O(n)
Función Copy (Copiar array multidimensional)	O(nm) n= filas, m=columnas

Tabla 1: Complejidad de las operaciones de la estructura de datos

Criterios de Diseño de la Estructura de Datos

- Se realiza la lectura del archivo csv con Pandas dataframe, para garantizar un cortado certero de los datos, y por las facilidades que tiene para rellenar los datos sin contestar con el valor string: “desconocido”, y de esta forma prevenir anomalías en la posterior construcción del árbol de decisión.

- El archivo se convierte a la estructura de array(arreglo) numpy, porque es mucho mas liviano para trabajar los datos, consume menor memoria y tenemos tamaños de datasets fijos.

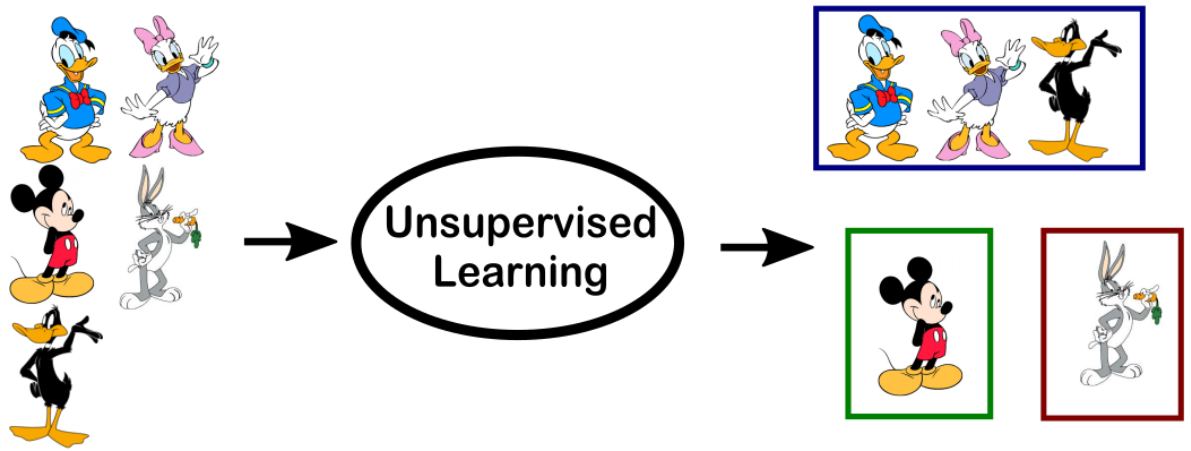
- La razón de usar esta estructura es la habilidad para llamar los datos por categorías como columnas y filas y que es más liviano que el panda’s dataframe, una opción similar pero mucho mas lenta.

Criterios de Diseño de la Estructura de Datos

- Se descarta el uso de ID3 por la existencia de atributos de tipo categórico de gran utilidad en el dataset, que el algoritmo es incapaz de procesar.
- Se decide usar el algoritmo C 4.5, y usar la medición de entropía de Shannon para la clasificación de los datos.
- Esta medida, aunque tarda un poco más que Gini la diferencia no suele ser mayor a un 2% en desempeño [*], además tiene una orientación mas adecuada para el "unsupervised learning" [*]

$$Entropy = -\sum_{j} p_j \log_2 p_j$$

Unsupervised Learning (Comparación)



ıi_educacionmad	ıi_ocupacionpadı	ıi_ocupacionmad	ıni_estratoviviendı	fami_nivelsisben	ımi_tieneinternet.	ı_tienecomputad	ı_trabajaactualmé	estu_antecedentes	estu_expectativas	punt_lenguaje	ɔunt_matematicas	punt_biologia	punt_quimica	punt_fisica
Educación pr…	Trabajador p…	Empleado de …	Estrato 3	Nivel 3	SI	SI	NO	NO	NO	45	47	45	41	37
Secundaria (…	Empleado obr…	Hogar	Estrato 2	Nivel 3	SI	SI	NO	NO	Desconocido	33	35	41	46	47
Secundaria (…	Otra activid…	Hogar	Estrato 2	Nivel 2	NO	SI	NO	NO	SI	38	49	46	53	41
Postgrado	Pensionado	Empleado de …	Estrato 5	No está clas…	SI	SI	NO	NO	Desconocido	39	46	49	35	42

Tal como en la imagen de ejemplo en nuestro dataset hay muchos datos con poca relación entre ellos. [*]

[*] Fuente de la imagen

Consumo de Tiempo y Memoria

	CONSTRUCCIÓN ARRAY	CONSTRUCCIÓN LIST
Conjunto de datos 1	0,002s	0.003s
Conjunto de datos 2	0,009s	0.014s
Conjunto de datos 3	0,011s	0,2s
Conjunto de datos 4	0,014s	0.018s
Conjunto de datos 5	0,021s	0.047s
Conjunto de datos 6	0,028s	0.069s
Conjunto de datos 7	0,037s	0.105s
Conjunto de datos 8	0,048s	0.125s
Conjunto de datos 9	0,068s	0.164s
Conjunto de datos 10	0,091s	0.197s

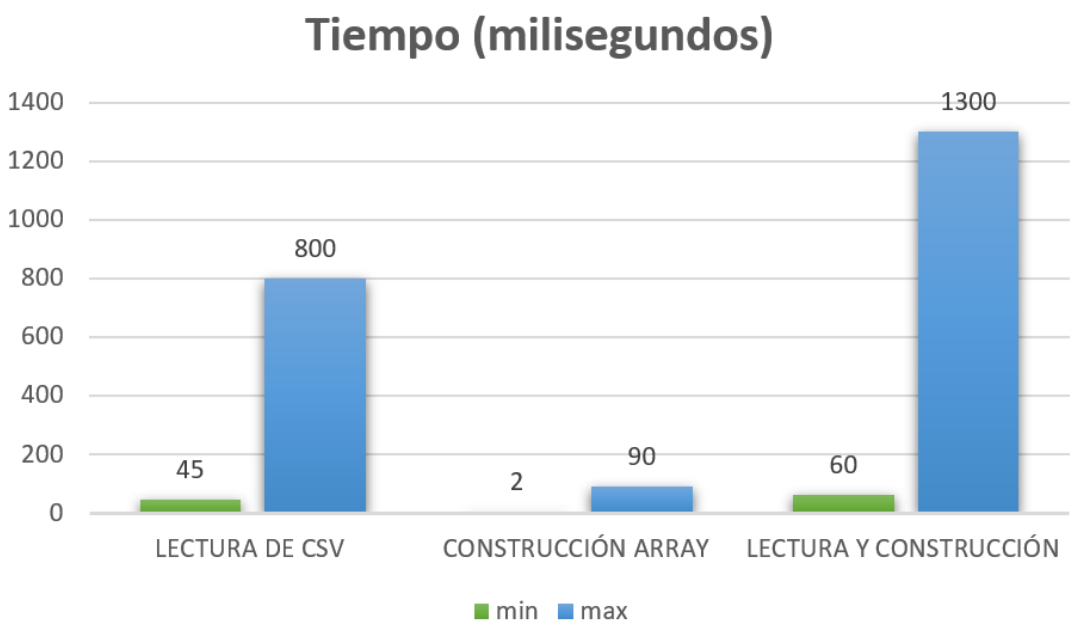
Tabla 3: Comparación de tiempos de ejecución de las operaciones de la estructura de datos entre el array Numpy y las Listas de Python

Comparación de tiempo de la construcción del archivo con arreglos de numpy vs. Con listas de python.

Consumo de Tiempo

	LECTURA DE CSV	CONSTRUCCIÓN ARRAY	LECTURA Y CONSTRUCCIÓN
Conjunto de datos 1	0,045s	0,002s	0,06s
Conjunto de datos 2	0,1s	0,009s	0,16s
Conjunto de datos 3	0,13s	0,011s	0,2s
Conjunto de datos 4	0,16s	0,014s	0,25s
Conjunto de datos 5	0,22s	0,021s	0,34s
Conjunto de datos 6	0,28s	0,028s	0,44s
Conjunto de datos 7	0,35s	0,037s	0,55s
Conjunto de datos 8	0,45s	0,048s	0,72s
Conjunto de datos 9	0,63s	0,068s	1s
Conjunto de datos 10	0,8s	0,091s	1,3s

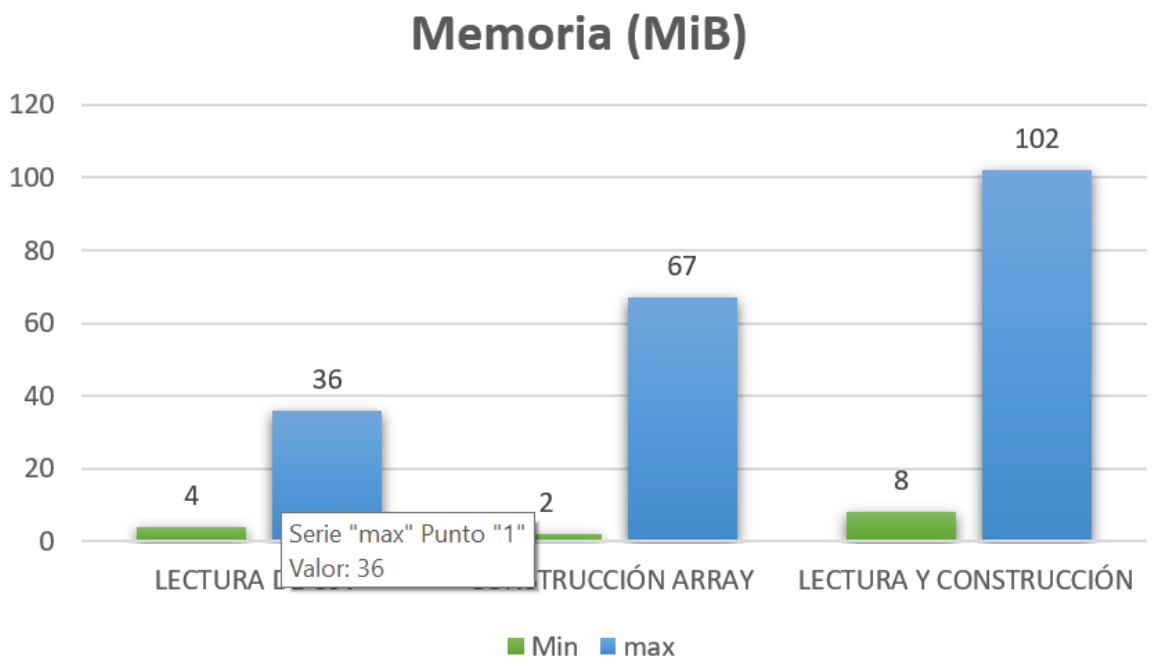
Tabla 2: Tiempos de ejecución de las operaciones de la estructura de datos con diferentes conjuntos de datos



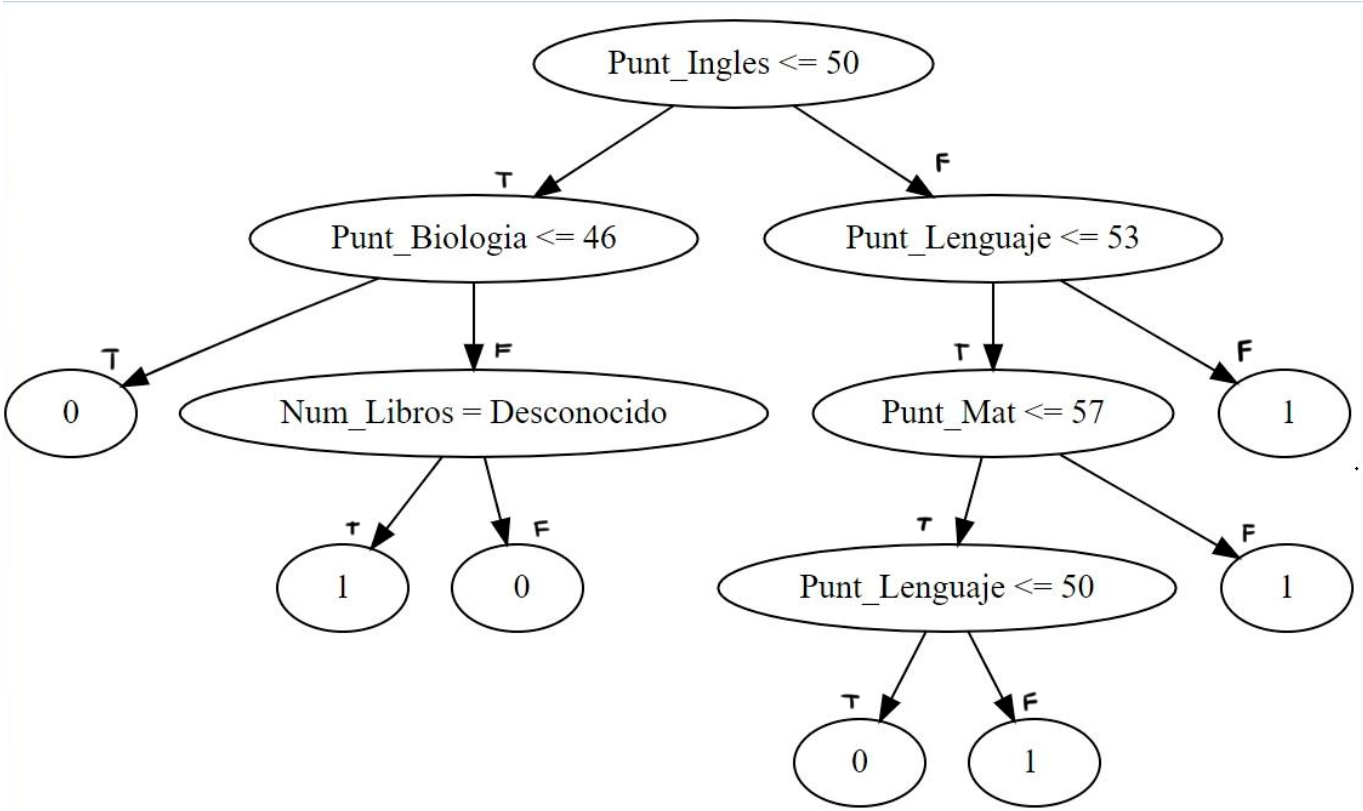
Consumo de Memoria

	LECTURA DE CSV	CONSTRUCCIÓN ARRAY	LECTURA Y CONSTRUCCIÓN
Conjunto de datos 1	4.0664 MiB	2.3007 MiB	8.2656 MiB
Conjunto de datos 2	4.0039 MiB	7.1445 MiB	17.2890 MiB
Conjunto de datos 3	7.0429 MiB	9.0585 MiB	15.1601 MiB
Conjunto de datos 4	4.1289 MiB	11.6757 MiB	16.7656 MiB
Conjunto de datos 5	7.0390 MiB	16.7851 MiB	24.6445 MiB
Conjunto de datos 6	13.4609 MiB	21.4257 MiB	34.9648 MiB
Conjunto de datos 7	16.7460 MiB	28.3164 MiB	42.2968 MiB
Conjunto de datos 8	20.3203 MiB	36.4960 MiB	55.0156 MiB
Conjunto de datos 9	28.1406 MiB	51.7148 MiB	77.8632 MiB
Conjunto de datos 10	35.6484 MiB	67.0234 MiB	101.8164 MiB

Tabla 4: Consumo de memoria de la estructura de datos con diferentes conjuntos de datos

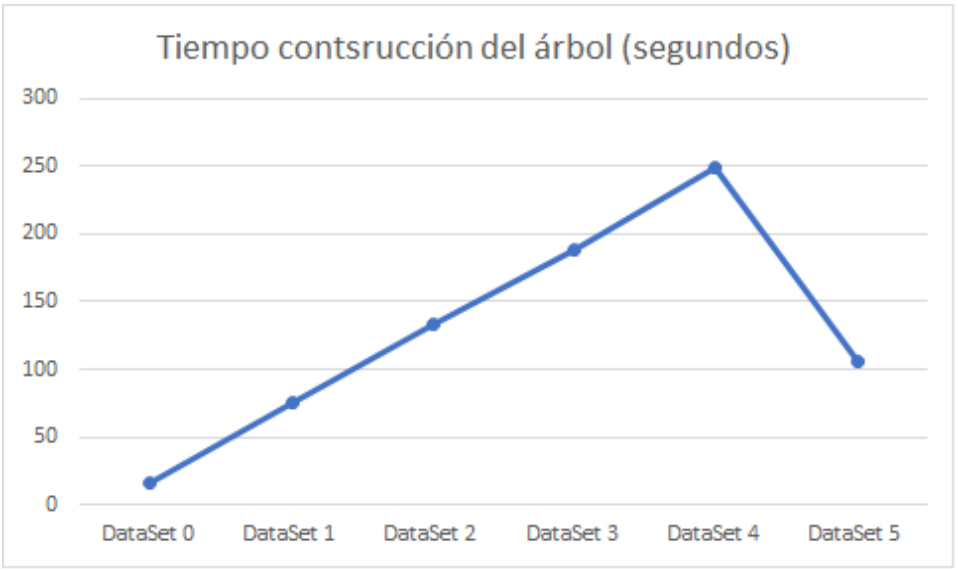


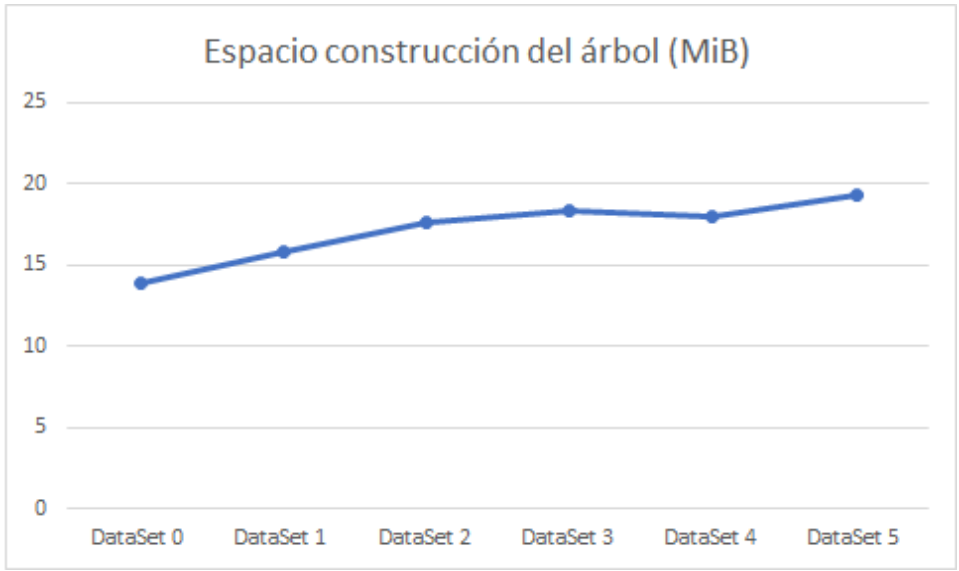
Software Desarrollado



Árbol obtenido con el algoritmo para el dataset-0.

Tiempo y espacio de construccion del arbol





Árboles Creados

```
{'punt_ingles <= 50.0': [{'punt_biologia <= 46.0': [0,
                                                    {'punt_quimica <= 48.0': [0,
                                                                              1]}]},
                         1]}
0.7546
```

Árbol de profundidad 3. Exactitud: 75%

```
{'punt_ingles <= 50.0': [{'punt_biologia <= 46.0': [0,
                                                    {'fami_numlibros = Desconocido': [1,
                                                                                      0]}]},
                         {'punt_lenguaje <= 53.0': [{'punt_matematicas <= 57.0': [{'punt_lenguaje <= 50.0': [0,
                                                                                                             1]},
                                                                                  1]},
                                                    1]}]}
0.76
```

Árbol de profundidad 4. Exactitud: 76%

```
{'punt_ingles <= 50.0': [{'punt_biologia <= 46.0': [{'fami_numlibros = Desconocido': [{'punt_lenguaje <= 47.0': [0,
                                                                                                                 {'punt_quimica <= 43.0': [0,
                                                                                                                                           1]}]},
                                                                                     0]},
                                                    {'punt_quimica <= 48.0': [{'fami_numlibros = Desconocido': [{'punt_lenguaje <= 47.0': [0,
                                                                                                                                           1]},
                                                                                                               0]},
                                                                              {'fami_numlibros = Desconocido': [1,
                                                                                                                {'punt_matematicas <= 53.0': [0,
                                                                                                                                              1]}]}]}]},
                         {'punt_lenguaje <= 53.0': [{'punt_matematicas <= 57.0': [{'punt_lenguaje <= 50.0': [{'punt_biologia <= 44.0': [0,
                                                                                                                                        1]},
                                                                                                             1]},
                                                                                  {'punt_ingles <= 60.0': [{'punt_biologia <= 43.0': [0,
                                                                                                                                      1]},
                                                                                                           1]}]},
                                                    {'punt_ingles <= 57.0': [{'punt_quimica <= 56.0': [{'punt_filosofia <= 42.0': [0,
                                                                                                                                   1]},
                                                                                                       1]},
                                                                             1]}]}]}
0.7732
```

Árbol de profundidad 5. >Exactitud: 77%

Comparación de diferentes datasets

```
--- DataSet 0 ---
{'punt_ingles <= 50.0': [{'punt_biologia <= 46.0': [0,
                                                    {'punt_quimica <= 48.0': [0,
                                                                              1]}]},
                         1]}
0.7546
--- DataSet 1 ---
{'punt_ingles <= 52.0': [{'punt_biologia <= 48.0': [0,
                                                    {'punt_quimica <= 51.0': [0,
                                                                              1]}]},
                         1]}
0.7569333333333333
--- DataSet 2 ---
{'punt_ingles <= 52.0': [{'punt_biologia <= 48.0': [0,
                                                    {'punt_quimica <= 51.0': [0,
                                                                              1]}]},
                         1]}
0.75508
```

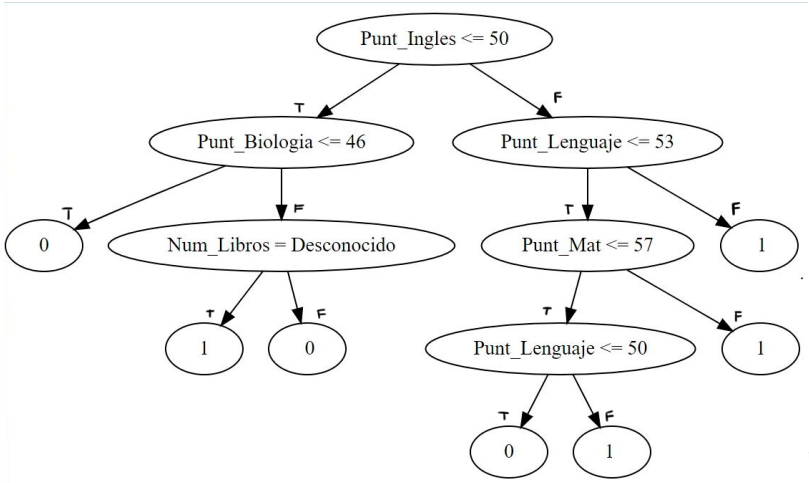
Comparación de diferentes datasets con la misma profundidad.
Se observa que dan resultados idénticos.

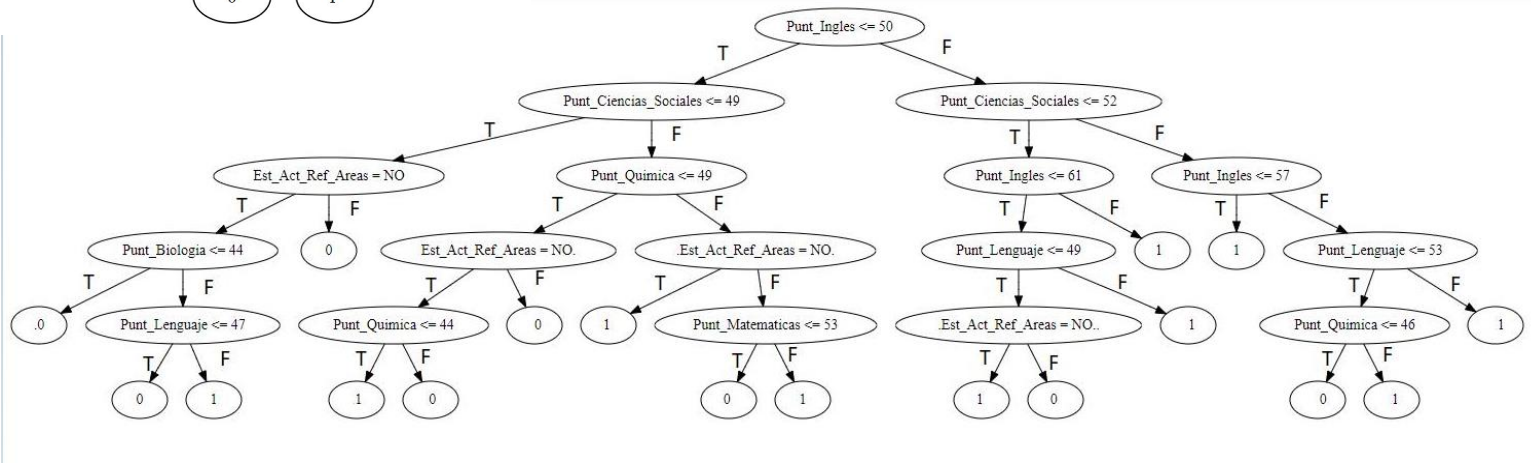
Comparación de árboles con menos datos categóricos

```
--- DataSet 0 - Depth 5 ---
{'punt_ingles <= 50.0': [{'punt_ciencias_sociales <= 49.0': [{'fami_numlibros = Desconocido': [{'punt_lenguaje <= 47.0': [0,
                                                                                                                          {'punt_biologia <= 44.0': [0,
                                                                                                                                                     1]}]},
                                                                                                                  0]},
                                                                                          {'punt_quimica <= 49.0': [{'fami_numlibros = Desconocido': [{'punt_lenguaje <= 48.0': [0,
                                                                                                                                                                                 1]},
                                                                                                                                                     0]},
                                                                                                                   {'fami_numlibros = Desconocido': [1,
                                                                                                                                                     {'punt_matematicas <= 53.0': [0,
                                                                                                                                                                                   1]}]}]}]},
                                                                 {'punt_ciencias_sociales <= 52.0': [{'punt_ingles <= 61.0': [{'punt_lenguaje <= 49.0': [{'fami_numlibros = Desconocido': [1,
                                                                                                                                                                                                           0]},
                                                                                                                                                                                  1]},
                                                                                                                                                     1]},
                                                                                                                             {'punt_ingles <= 57.0': [1,
                                                                                                                                                      {'punt_lenguaje <= 53.0': [{'punt_quimica <= 46.0': [0,
                                                                                                                                                                                                           1]},
                                                                                                                                                                                  1]}]}]}]}
0.7896

--- DataSet 0 - Depth 5 ---
{'punt_ingles <= 50.0': [{'punt_ciencias_sociales <= 49.0': [{'estu_actividadrefuerzoareas = NO': [{'punt_biologia <= 44.0': [0,
                                                                                                                             {'punt_lenguaje <= 47.0': [0,
                                                                                                                                                        1]}]},
                                                                                                                     0]},
                                                                                             {'punt_quimica <= 49.0': [{'estu_actividadrefuerzoareas = NO': [{'punt_quimica <= 44.0': [0,
                                                                                                                                                                                      1]},
                                                                                                                                                        0]},
                                                                                                                      {'estu_actividadrefuerzoareas = NO': [1,
                                                                                                                                                            {'punt_matematicas <= 53.0': [0,
                                                                                                                                                                                          1]}]}]}]},
                                                                 {'punt_ciencias_sociales <= 52.0': [{'punt_ingles <= 61.0': [{'punt_lenguaje <= 49.0': [{'estu_actividadrefuerzoareas = NO': [1,
                                                                                                                                                                                                              0]},
                                                                                                                                                                                  1]},
                                                                                                                                                     1]},
                                                                                                                             {'punt_ingles <= 57.0': [1,
                                                                                                                                                      {'punt_lenguaje <= 53.0': [{'punt_quimica <= 46.0': [0,
                                                                                                                                                                                                           1]},
                                                                                                                                                                                  1]}]}]}]}
0.7886
```

Comparación de árboles con menos datos categóricos





Comparación de árboles con datos categóricos

Como se puede apreciar en la diapositiva anterior, al eliminar las columnas correspondientes a "Numero de libros leídos" y "Tiene Etnia", además de las relacionadas a los posibles cursos previos tomados se mejora el árbol evitando datos de tipo "Desconocido".

Aunque estos datos puedan ser útiles en teoría la gran cantidad de "vacíos" que tienen afectan el árbol negativamente.

```
[{'estu_tieneetnia = Desconocido': [{'punt_lenguaje <= 47.0': [0,
```

```
[{'fami_numlibros = Desconocido': [{'punt_lenguaje <= 47.0': [0,
```

Complejidad

Método	Complejidad
Train_test_split:	O(n)
Check_purity:	O(n)
Create_leaf:	O(n)
Get_potential_splits:	O(nxm)
Calculate_entropy:	O(n)
Calculate_mse:	O(n)
Calculate_overall_metric:	O(n)
Determine_Best_Split:	O(nxm)
Split_Data:	O(n)
Determine_Type_Of_Feature:	O(nxm)
Decision_Tree_Algorithm:	O(n^2xm^2)
Predict_Example:	O(1)
Make_Prediction:	O(n)
Calculate_Accuracy:	O(n)
total	O(n^2xm^2)
* n es la cantidad de filas del dataset	
* m es la cantida de columnas del arreglo	

Estructura del programa

Train_test_split: En caso de hacer una prueba rápida construye un test dataset con algunos elementos del train dataset.
Check_purity: Devuelve Si/No si los datos están mezclados.
Create_leaf: Crea un nodo hoja.
Get_potential_splits: Devuelve un diccionario con los puntos de corte potenciales.
Calculate_entropy: Calcula la entropía de un grupo determinado.
Calculate_mse: Calcula el "Mean squared error". Siempre es positivo, mientras más cerca a cero mejor.
Calculate_overall_metric:Calcula métricas para los grupos que se crean por un nuevo split.
Determine_Best_Split: Escoge las condiciones para hacer el corte/split.
Split_Data: Corta el grupo en dos subgrupos: superior e inferior.
Determine_Type_Of_Feature: Determina si se está tratando una variable numérica o categórica.
Decision_Tree_Algorithm: Método recursivo. Crea el árbol y es quien llama los otros métodos.
Predict_Example: Método de predicción con un solo ejemplo.
Make_Prediction: Método de predicción con todo un dataset.
Calculate_Accuracy: Calcula la exactitud de la predicción versus el árbol creado.