**CLASIFICACIÓN BASE DE DATOS DE CRITICAS PELICULAS**

**Reprocesamiento Data set**

Para poder aplicar las técnicas de clasificación como Arboles, Método de Bayes y K-nearest primero se realizó un trabajo extenso en el procesamiento de los documentos, para tener el datase listo para su respectiva clasificación.

En nuestra base de datos se utiliza un conjunto de datos divididos en críticas negativas y positivas de películas con el fin de que existan atributos válidos para una clasificación se realiza una reprocesamiento con los siguientes pasos:

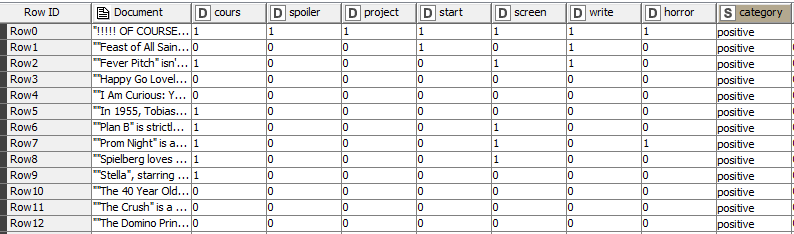
1. Se agrega una columna “category” donde se pone si la crítica pertenece al conjunto positivo o negativo.
2. Eliminación de puntos del documento
3. Eliminar los números del documento.
4. Seleccionar las palabras con más significado para la crítica, en este caso se seleccionan solo los sustantivos y adjetivos en ingles de los conjuntos de datos, y se eliminan los adverbios y conjunciones del documento.

Por Ejemplo, la palabra “This is an antology horror film” nos quedaría “antology horror film”. Como se puede ver los artículos “is”, “an”.

1. Se aplica un algoritmo de le matización para eliminar inflexiones en las palabras y convertirlas a su raíz, por ejemplo:

|  |  |  |
| --- | --- | --- |
| Palabra | Regla | Resultado |
| Conflated | AT to ATE | Conflate |
| Trouble | BL to BlE | Trouble |
| Sized | IZ to IZE | Size |

1. Se aplica un “Bag Of Words” el cual crea una columna en nuestra tabla de asociación con cada palabra contenidas en un documento.
2. Se realiza cuenta la frecuencia de una palabra en cada documento, esto con el objetivo de asociar que palabras se repiten más en nuestro subconjunto de críticas positivas y negativas.
3. Finalmente se obtiene la tabla siguiente:



La cual nos muestra los documentos los cuales son las críticas que están en la base de datos, palabras las cuales se muestran el número de veces que aparecen en cada crítica o documento y finalmente la categoría, la cual puede ser negativa o positiva dependiendo del documento dado por los datos de prueba.

Todo este pre procesamiento se realiza con el objetivo de poder realizar una clasificación por las palabras más frecuentes en un documento, y poder saber con precisión si esta palabra tiene una influencia positiva o negativa en las críticas.

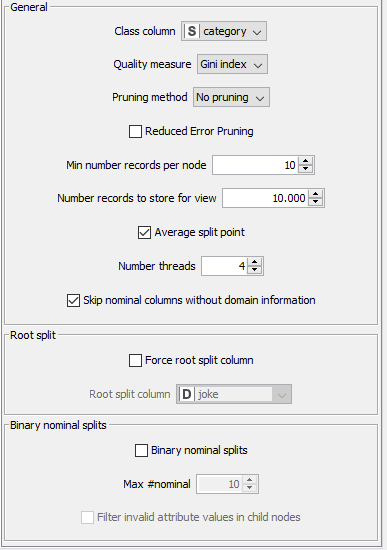
**Clasificación Por Arboles**

Después de procesar nuestro conjunto de pruebas, realizamos una clasificación mediante un árbol binario. Nuestro modelo es construido a partir de las palabras de mayor frecuencia en cada documento y la categoría de las críticas, si son negativas o positivas. El objetivo de nuestro árbol es que clasifique según las palabras si una crítica es positiva o negativa.

**Modelo de árbol**

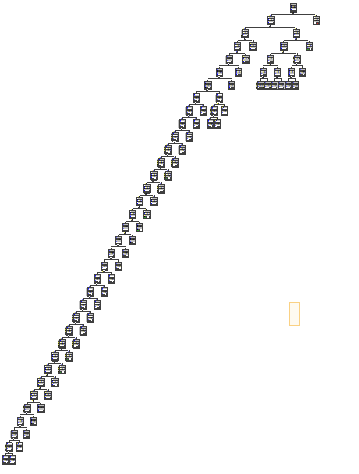
1. **Primera configuración**
   1. **Parámetros**

Para esta primera configuración se colocan los siguientes parámetros.



**1.2 Modelo de árbol**

Como podemos ver en la siguiente imagen el modelo de Árbol tiene una gran profundidad



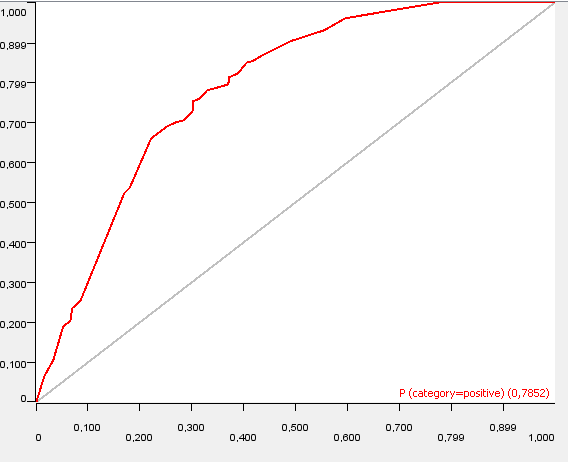
**1.3 Matriz de Confusión**

Los resultados de la matriz de confusión fueron los siguientes:

|  |  |  |
| --- | --- | --- |
| **Variables** | **Positivo** | **Negativo** |
| **Positivo** | 228 | 72 |
| **Negativo** | 95 | 205 |

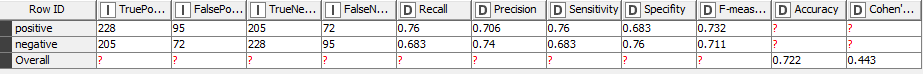
**1.4 Curva RoC:**

Como se puede ver en la siguiente imagen la curva para la clasificación de la categoría de la crítica está por encima de la recta diagonal lo cual es positivo.



**1.5 Estadísticas de precisión:**

En la siguiente tabla se muestra variables para la medición de precisión del modelo de Arboles:

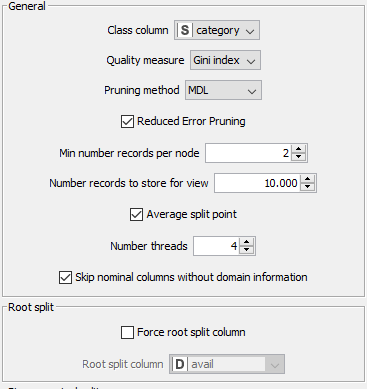


Las variables “Acuracy” es igual a 0.706 y “Recall” es de 0.76.

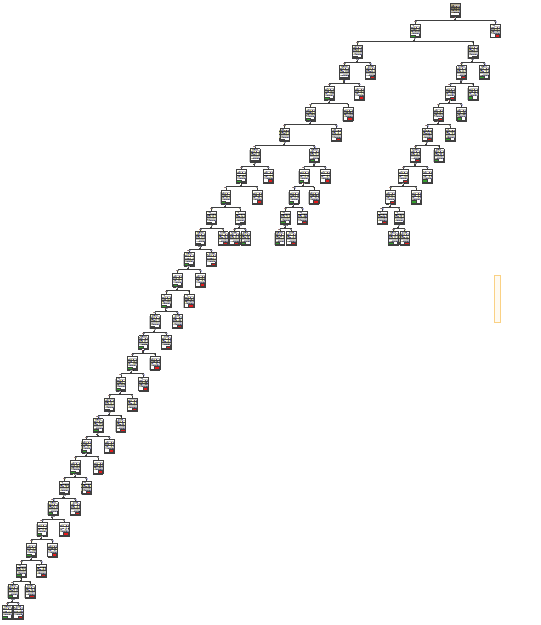
1. **Segunda Configuración**

**2.1 Parámetros:**

Para esta configuración se tiene los parámetros en la siguiente imagen, cabe destacar que en esta configuración se utiliza el método de poda para el árbol, los parámetros principales que se cambiaron fue el número de records por nodo y se activó el método de poda.



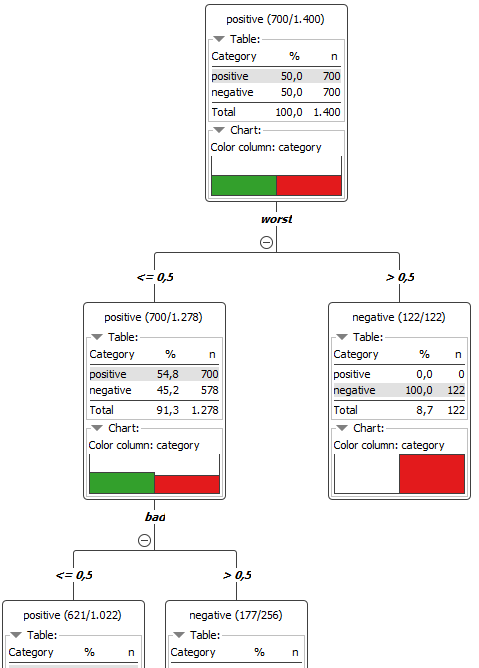
**2.2:** **Modelo de Árbol:** El resultado del modelo de árbol hecho con el algoritmo de poda fue el siguiente:



La imagen anterior nos muestra un árbol binario donde los nodos son las palabras, y dependiendo de la frecuencia de aparición de cada palabra el árbol la clasifica en malo o bueno.

Si bien el árbol resultante es profundo, cabe destacar que existen una gran cantidad de combinaciones de palabras para cada documento.

Sin embargo, como se muestra en la siguiente imagen, vemos que la palabra “worst” que es nuestro nodo raíz y en ingles una palabra negativa, cuando existe una frecuencia de esta palabra mayor a 0.5, de inmediato clasifica la crítica en una categoría Negativa, lo que comprueba que el modelo está funcionando correctamente:



**2.3 La matriz de confusión:**

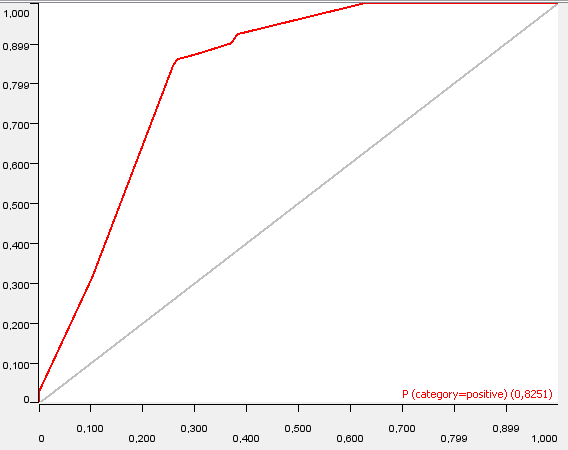
Esta matriz nos muestra la cantidad de datos que nuestro modelo de clasificación pudo predecir.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Positivo** | **Negativo** |
| **Positivo** | 258 | 42 |
| **Negativo** | 40 | 220 |

Como se muestra en la tabla anterior fueron 258 verdaderos positivos, 42 falsos positivos 40 verdaderos negativos y 220 falsos negativos. Lo cual muestra un gran acierto a la hora de predecir categorías por parte de nuestro árbol.

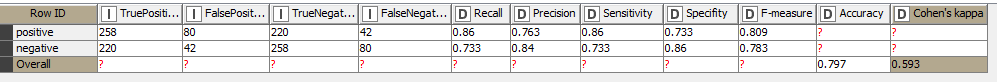
**2.4 Curva RoC:**

Como podemos ver la curva de ROC en cuanto a la hora de clasificar categorías negativas, como podemos observar la curva está por encima de la recta diagonal, con lo cual nos indica un resultado bueno, y con el área bajo la curva de 0.8251 que está muy cerca de 1.



* 1. **Estadísticas de precisión:**

Se muestra en la siguiente imagen

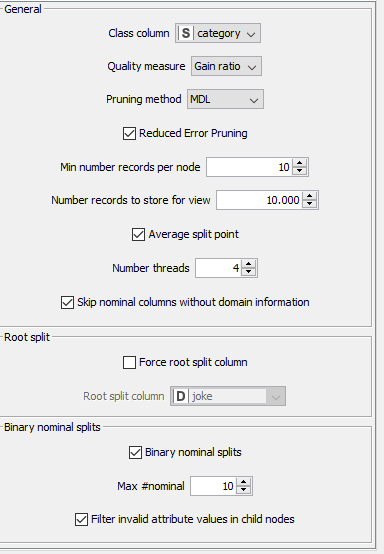


Para destacar, se encuentra que la precisión es de 0.763 y el Recall es de 0.86 lo que significa que los falsos positivos y verdaderos negativos son menores, respectivamente.

**3 Tercera configuración**

**3.1 Parámetros**

Para esta prueba se elige calcular los atributos nominales de una forma binaria seleccionando la opción “Binary Nominal Splits”. Esto con el objetivo de mejorar la precisión de la predicción de nuestro modelo. Además para reducir la complejidad de estos cálculos ya que la forma de selección binaria es más precisa pero más compleja se selecciona un conjunto de rangos para que sea más rápido el cálculo, en este caso fue de 10.



3.2 **Modelo Árbol Binario**

El modelo del árbol resultante fue el siguiente, se destaca que la raíz del Árbol otra vez se repite la palabra “Worst “como en las tres anteriores clasificaciones:



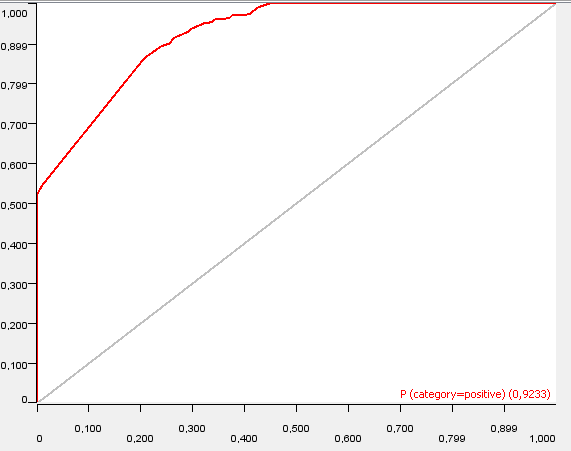
3.3 **Matriz de confusión**

Esta matriz nos muestra la cantidad de datos que nuestro modelo de clasificación pudo predecir.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Positivo** | **Negativo** |
| **Positivo** | 259 | 41 |
| **Negativo** | 32 | 238 |

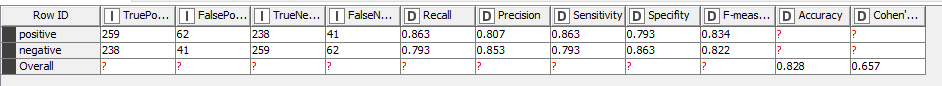
**3.4 Curva ROC**

La siguiente imagen muestra la curva ROC con un área bajo la curva de 0.9233



3.5 **Estadísticas de precisión**

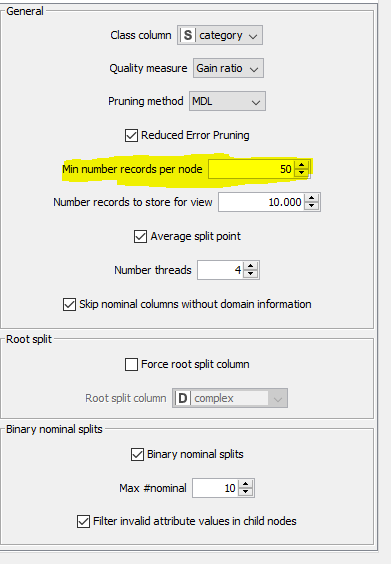
Para el Modelo se obtienen los siguientes datos sobre su precisión.



**4. Cuarta configuración**

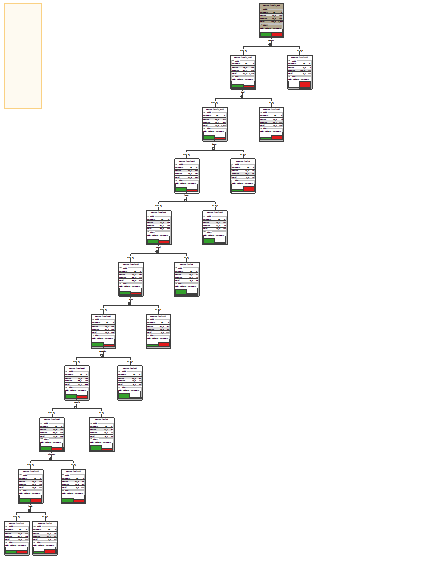
**4.1 Parámetros:**

Para esta configuración se incrementa el número mínimo de archivos por nodo a 50, con el objetivo de generalizar aún más la clasificación y que el árbol resultante no sea tan profundo.



4.2 **Modelo Árbol Binario:**

Como se muestra en la siguiente imagen el arbol resultante es mucho menos profundo que en cualquiera de las anteriores configuraciones.

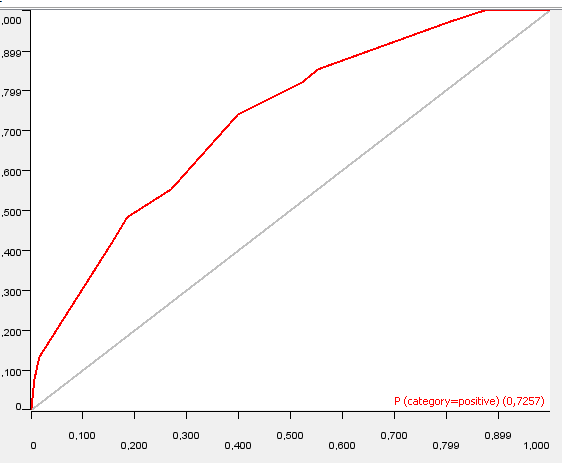


**4.3 Matriz de confusión**

|  |  |  |
| --- | --- | --- |
| **Variables** | **Positivo** | **Negativo** |
| **Positivo** | 222 | 78 |
| **Negativo** | 120 | 180 |

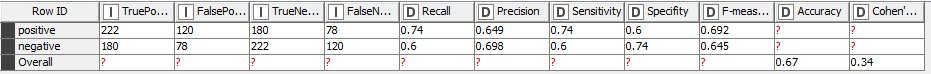
**4.4 Curva ROC**

La grafica muestra una curva positiva por encima de la diagonal con un área bajo la curva de 0.7257



**4.5 Estadísticas de precisión**

La precisión del modelo, como se puede ver en la siguiente tabla es de 0.67 un poco más de la mitad.



**Comparación Atributos Diferentes Configuraciones**

En la siguiente tabla se muestra la comparación entre las diferentes configuraciones para nuestro modelo de clasificación

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Configuración | Área Bajo la curva ROC | Accuracy | Cohen’s Cappa | TruePositive | | FalsePositive | | TrueNegative | | FalseNegatie | |
| P | N | P | N | P | N | P | N |
| Configuración 1 | 0.7852 | 0.722 | 0.443 | 228 | 205 | 95 | 72 | 205 | 228 | 72 | 95 |
| Configuración 2 | 0.8251 | 0.797 | 0.593 | 258 | 220 | 80 | 42 | 220 | 258 | 42 | 80 |
| Configuración 3 | **0.9233** | **0.822** | **0.657** | **259** | **238** | **62** | **41** | **238** | **259** | **41** | **62** |
| Configuración 4 | 0.7257 | 0.670 | 0.340 | 222 | 180 | 120 | 78 | 180 | 222 | 78 | 120 |

**Conclusión**

Como se puede ver en la tabla el modelo que tuvo una mejor tasa de clasificación y precisión fue el número tres el cual se utilizó la forma de clasificar “Binary Nominal Splits”. La cual tiene una precisión de 82% y un Coeficiente de Cohen de 0.657. Además, tiene más aciertos True Positive en las categorías Positivas(P) y Negativa(N).

La configuración 4 como podemos observar en la tabla fue la que peores resultados obtuvo, con una precisión de 0.670, y con un número menor de True Positive y True Negative. Esto se debe al incremento de archivos por nodo en la configuración inicial , y pese a que es un árbol de menor tamaño como se mostró , en este caso esto lleva a que la clasificación sea menos precisa.