UNIVERSIDAD DE CORDOBA

FACULTAD DE INGENIERIA

INGENIERIA DE SISTEMAS

TALLER CLASIFICACIÓN CON ÁRBOLES DE DECISIÓN Y BAYES EN KNIME

ELECTIVA APRENDIZAJE AUTOMATICO

INTEGRANTES:

DEIMER YAMIT HERNANDEZ DE LA ROSA

DEYSER OROZCO YEPES

PROFESOR:

OSWALDO VELEZ LANGS

2020

1. Compresión de datos

|  |  |  |
| --- | --- | --- |
| Variable | Descripción | Tipo de dato |
| |  | | --- | | word\_make | | |  | | --- | | word\_make | | String |
| |  | | --- | | word\_address | | |  | | --- | | word\_address | | String |
| |  | | --- | | word\_all | | |  | | --- | | word\_all | | String |
| |  | | --- | | word\_3d | | |  | | --- | | word\_3d | | String |
| |  | | --- | | word\_our | | |  | | --- | | word\_our | | String |
| |  | | --- | | word\_over | | |  | | --- | | word\_over | | String |
| |  | | --- | | word\_remove | | |  | | --- | | word\_remove | | String |
| |  | | --- | | word\_internet | | |  | | --- | | word\_internet | | String |
| |  | | --- | | word\_order | | |  | | --- | | word\_order | | String |
| |  | | --- | | word\_mail | | |  | | --- | | word\_mail | | String |
| |  | | --- | | word\_receive | | |  | | --- | | word\_receive | | String |
| |  | | --- | | word\_will | | |  | | --- | | word\_will | | String |
| |  | | --- | | word\_people | | |  |  | | --- | --- | | word\_people |  | | String |
| |  | | --- | | word\_report | | |  | | --- | | word\_report | | String |
| |  | | --- | | word\_addresses | | |  | | --- | | word\_addresses | | String |
| word\_free | word\_free | String |
| |  | | --- | | word\_business | | |  | | --- | | word\_business | | String |
| |  | | --- | | word\_email | | |  | | --- | | word\_email | | String |
| |  | | --- | | word\_you | | |  | | --- | | word\_you | | String |
| |  | | --- | | word\_credit | | |  | | --- | | word\_credit | | String |
| |  | | --- | | word\_your | | |  | | --- | | word\_your | | String |
| |  | | --- | | word\_font | | |  | | --- | | word\_font | | String |
| |  | | --- | | word\_000 | | |  | | --- | | word\_000 | | String |
| |  | | --- | | word\_money | | |  | | --- | | word\_money | | String |
| |  | | --- | | word\_hp | | |  | | --- | | word\_hp | | String |
| |  | | --- | | word\_hpl | | |  | | --- | | word\_hpl | | String |
| |  | | --- | | word\_george | | |  | | --- | | word\_george | | String |
| |  | | --- | | word\_650 | | |  | | --- | | word\_650 | | String |
| |  | | --- | | word\_lab | | |  | | --- | | word\_lab | | String |
| |  | | --- | | word\_labs | | |  | | --- | | word\_labs | | String |
| |  | | --- | | word\_telnet | | |  | | --- | | word\_telnet | | String |
| |  | | --- | | word\_857 | | |  | | --- | | word\_857 | | String |
| |  | | --- | | word\_data | | |  | | --- | | word\_data | | String |
| |  | | --- | | word\_415 | | |  | | --- | | word\_415 | | String |
| |  | | --- | | word\_85 | | |  | | --- | | word\_85 | | String |
| |  | | --- | | word\_technology | | |  | | --- | | word\_technology | | String |
| |  | | --- | | word\_1999 | | |  | | --- | | word\_1999 | | String |
| |  | | --- | | word\_parts | | |  | | --- | | word\_parts | | String |
| |  | | --- | | word\_pm | | |  | | --- | | word\_pm | | String |
| |  | | --- | | word\_direct | | |  | | --- | | word\_direct | | String |
| |  | | --- | | word\_cs | | |  | | --- | | word\_cs | | String |
| |  | | --- | | word\_meeting | | |  | | --- | | word\_meeting | | String |
| |  | | --- | | word\_original | | |  | | --- | | word\_original | | String |
| |  | | --- | | word\_project | | |  | | --- | | word\_project | | String |
| |  | | --- | | word\_re | | |  | | --- | | word\_re | | String |
| |  | | --- | | word\_edu | | |  | | --- | | word\_edu | | String |
| |  | | --- | | word\_table | | |  | | --- | | word\_table | | String |
| |  | | --- | | word\_conference | | |  | | --- | | word\_conference | | String |

2) Al implementar el nodo Statistics se obtuvieron los siguientes histogramas, se observó que los datos son asimétricos a la izquierda, esto quiere decir que en el modelo la mayor parte de los datos tienen el valor de nonspam. Y son pocos los datos con spam, en los diagramas tampoco se encontraron valores atípicos o fuera de rango.

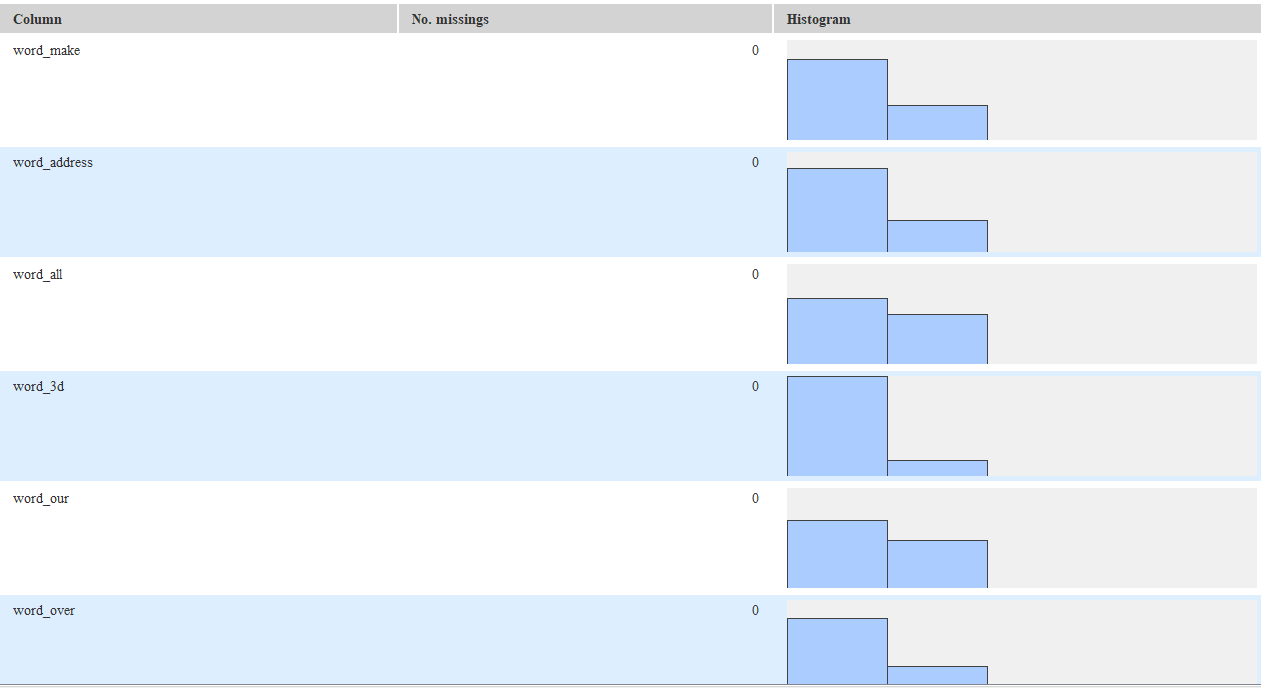


Ilustración 1

3)

4)

Al Implementar nodo “Decision tree learner (deprecated)”, se observó que el atributo que más importa es el de la palabra Word\_remove ya que este, clasifica mejor las palabras cuando toma el valor de verdadero (1), esto se puede ver en la ilustración 4, sin embargo, el modelo sigue siendo complejo, debido a que muestra un árbol gigantesco (ilustración 3).

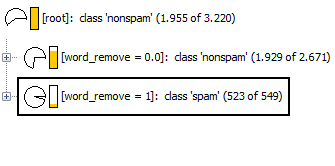


Ilustración 2

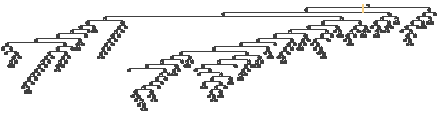


Ilustración 3 modelo del árbol obtenido

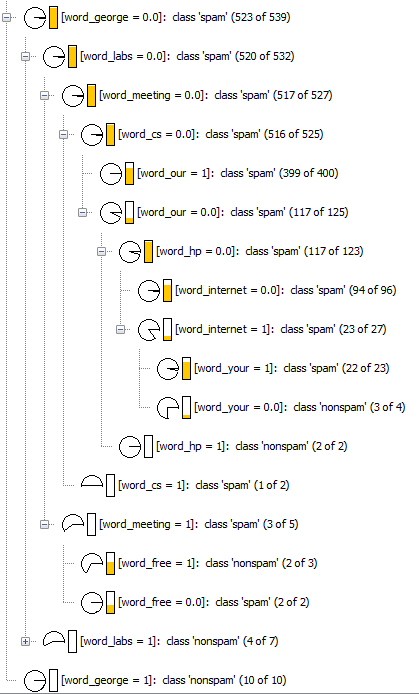


Ilustración 4

5) Modelado Predicción. Desplegar nodo “Decision tree predictor (deprecated)”, desplegar “Árbol” interpretar y comparar los resultados con la clasificación del anterior modelo.

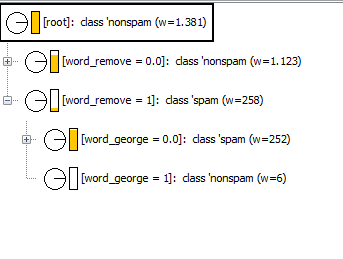


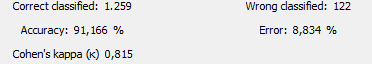
Ilustración 5

En comparación con los resultados obtenidos del modelo anterior se observa que, en el modelo predicho en el árbol, el atributo que sigue teniendo relevancia es Word\_remove, pero en ambos casos el árbol sigue siendo gigantesco.

6) Para la evaluación del modelo desplegar nodo “Scorer (Deprecated)”; analizar e interpretar los resultados de la evaluación.

Resultados:





Como se puede observar el modelo se clasifican correctamente 1.259 del total de los datos y arroja un porcentaje de accuracy del 91,166%, y el margen de error es de 8,834%, lo cual indica que los datos se están prediciendo correctamente a pesar de la complejidad del modelo.

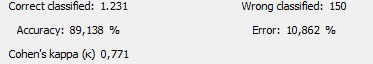
Realizar 3 pruebas cambiando los parámetros del “árbol de Aprendizaje”: Medidas de calidad, método de poda, y demás parámetros que permitan seleccionar el modelo de mayor confiabilidad.

Prueba # 1

Resultados:

|  |  |
| --- | --- |
| Parámetros |  |
| Medidas de calidad | Gain index |
| Método de poda | pml |
| número de hilos | 5 |
| número mínimo de registros por nodo | 10 |



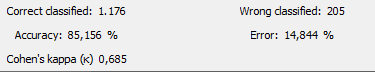


Prueba # 2

Resultados:

|  |  |
| --- | --- |
| Parámetros |  |
| Medidas de calidad | Gain ratio |
| Método de poda | Pml |
| número de hilos | 3 |
| número mínimo de registros por nodo | 100 |



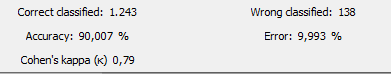


Prueba # 3

Resultados:

|  |  |
| --- | --- |
| Parámetros |  |
| Medidas de calidad | Gain ratio |
| Método de poda | PML |
| número de hilos | 1 |
| número mínimo de registros por nodo | 2 |





De los resultados de las pruebas el modelo de mayor confiabilidad fue el de la prueba numero tres, ya que tuvo el menor porcentaje de error de las tres pruebas y cabe decir que entre mayor sea el numero minino de registros por nodo, mayor serán los porcentajes de error, por lo tanto, el modelo no será tan eficiente para predecir los resultados y esto se puede comprobar con los porcentajes de accuracy debido a que en ambos casos(prueba1#, prueba#2) estos disminuyeron.