# Temas Selectos de Inteligencia Artificial

## Reporte Clasificación de Textos

Edgar Adrián Nava Romo

Profesor: Dr. Hiram Calvo

12 de Junio de 2018

# Índice

**Introducción**
Se tiene una base de datos de un grupo de 201 universitarios y profesores divididos en 101 hombres y 100 mujeres donde cada uno escribió su biografía con sus propias palabras,

**Objetivo**
Estos datos fueron recabados con el objetivo de encontrar una relación entre el lenguaje utilizado por los estudiantes para ver qué tipo de apego tienen más, ya sea si son Seguros, Preocupados, Evitantes o Temerosos.

Se utilizará un software clasificador de datos llamado Weka en su versión 3.2.8 y diferentes algoritmos de clasificación de texto de forma supervisada, esto es, indicando las clase en la que pertenece cada biografía, utilizando diferente número de atributos de acuerdo a la importancia que tengan en general, se intentará encontrar algún patrón entre los alumnos y profesores que ayude a facilitar la clasificación de alumnos para tener un mejor control o comprobar que el ser humano tiende a usar un mismo lenguaje cuando se encuentra en ciertas circunstancias.

**Desarrollo experimental**
Dado que la base de datos estaba en un archivo .txt, fue necesario convertirlo en formato ARFF para que el software utilizado pudiera leer los datos adecuadamente. Para esto, se agregó una relación al principio del documento (@RELATION) llamado “APEGO” y cinco atributos (@ATTRIBUTES) que son las variables que se encontrará el software a la hora de la clasificación, estás fueron:

@ATTRIBUTE numero string
Este es el ID del sujeto, se omitirá en un futuro para que no afecte en los resultados pero es importante etiquetar a cada sujeto para tener un mejor control sobre los datos.

@ATTRIBUTE Genero {1,2}
Es bien sabido que el lenguaje entre hombres y mujeres suele cambiar y al mismo tiempo tener coincidencias entre su mismo género, es por eso que se agregó este atributo, ayudando así a que los resultados mejoren.

@ATTRIBUTE Edad {18,19,20,21,22,23,24,25,26,27,28,29,31,37,40,41}

Se agregó la edad como atributo para encontrar si hay una relación crucial con el tipo de apego y la edad que tiene cada estudiante y profesor.

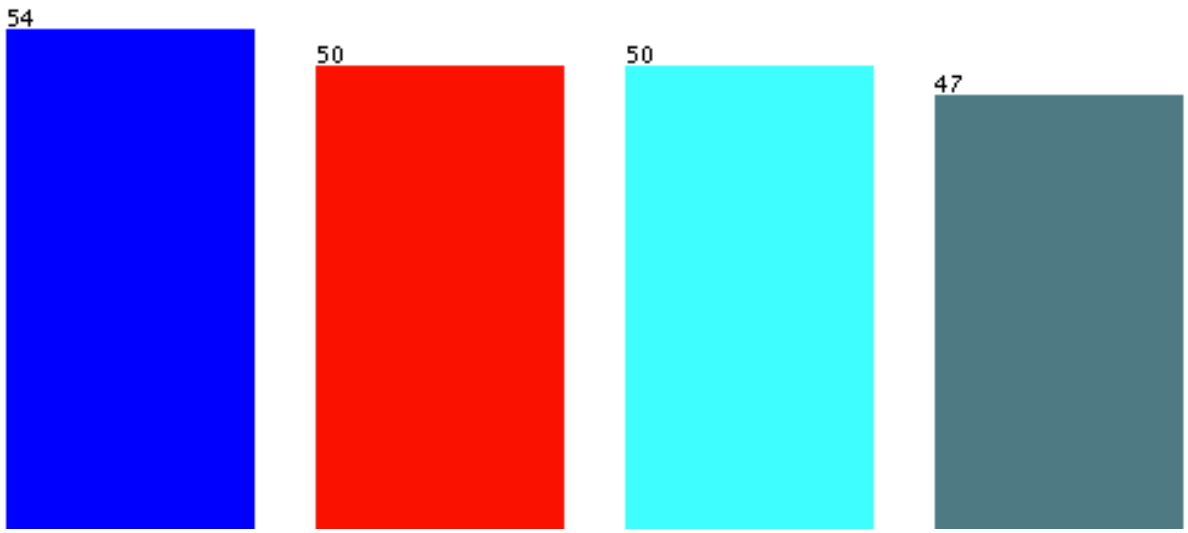
@ATTRIBUTE Class {T,E,S,P}
Dado que será una clasificación de texto supervisada, es importante tener un atributo con todas las clases que se tomarán en cuenta, en este caso se utilizaron como variables las iniciales de cada tipo de apego a clasificar: Temeroso, Evitante, Seguro y Preocupado.

@ATTRIBUTE Temp string
Contiene todas las biografías de alumnos y profesores, se trabajará mayormente sobre este atributo.

Luego de agregar los atributos, se agregaron los datos con @DATA separados por comas o comillas (en caso de que el tipo de dato sea una cadena declarada).

Una vez listo el archivo .Arff, se cargó al software y se encontró que los datos no estaban equilibrados, sin embargo el archivo generado fue usado como referencia para ver si los cambios a realizar repercutirían de forma positiva o negativa.

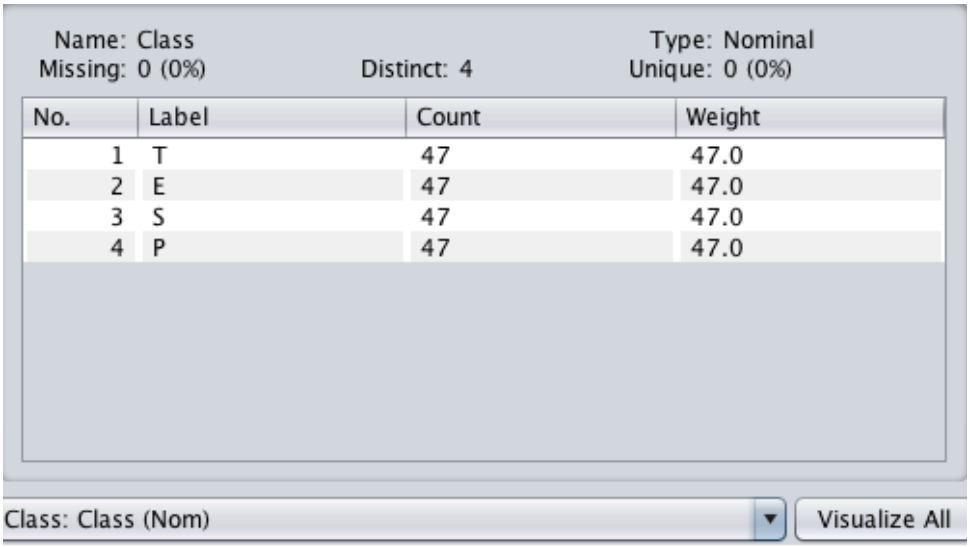
| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | T | 54 | 54.0 |
| 2 | E | 50 | 50.0 |
| 3 | S | 50 | 50.0 |
| 4 | P | 47 | 47.0 |

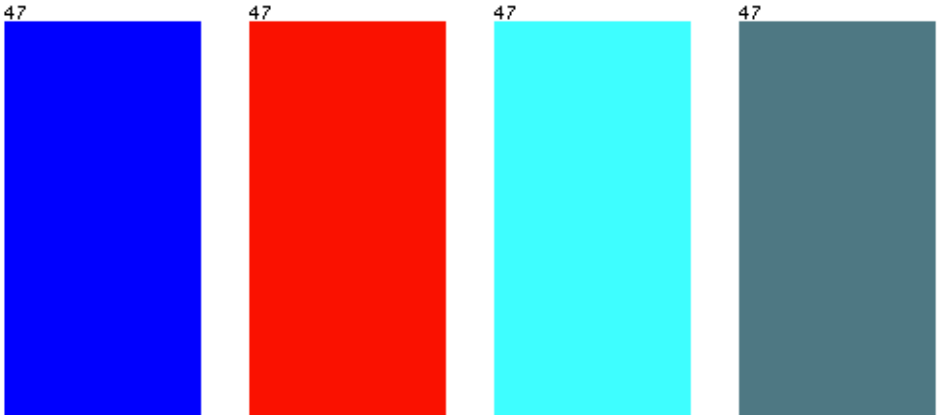


Para empezar a experimentar con este archivo usando el filtro StringToWordVector, que se encarga de separar los atributos de @DATA para facilitar el manejo de datos este filtro encontró 1,287 atributos.

Por otro lado, se equilibraron los datos de dos formas diferentes, usando un filtro llamado “ClassBalancer” el cual no sirvió ya que cortó de forma decimal los datos y fue prácticamente inservible.

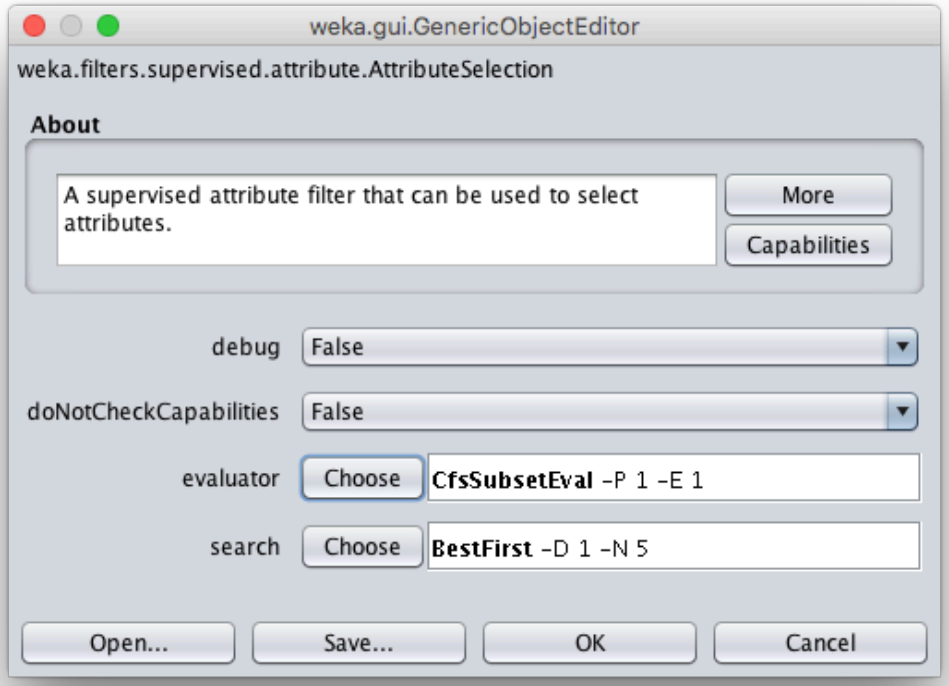
El segundo método fue manual, se equilibró todo de acuerdo al género y la clase quedando en 94 hombres y 94 mujeres con 47 biografías de cada una de las 4 clases, dando un total de 188 instancias.
En este segundo archivo se hizo el mismo proceso que en el primero usando “StringToWordVector” y encontró un total de 5718 atributos

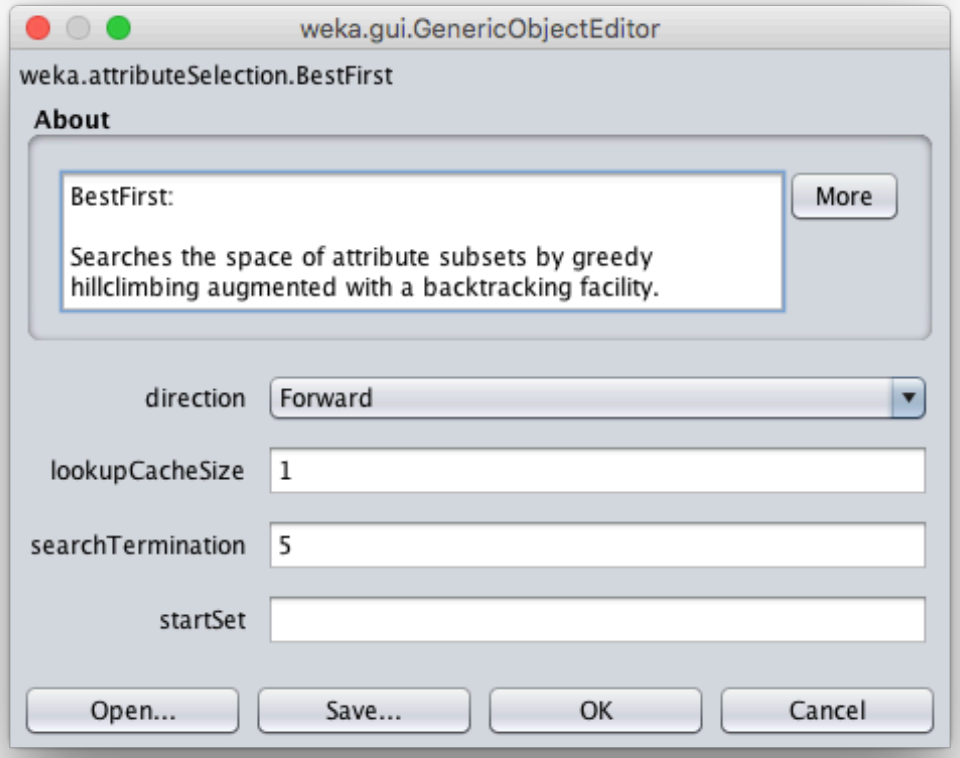


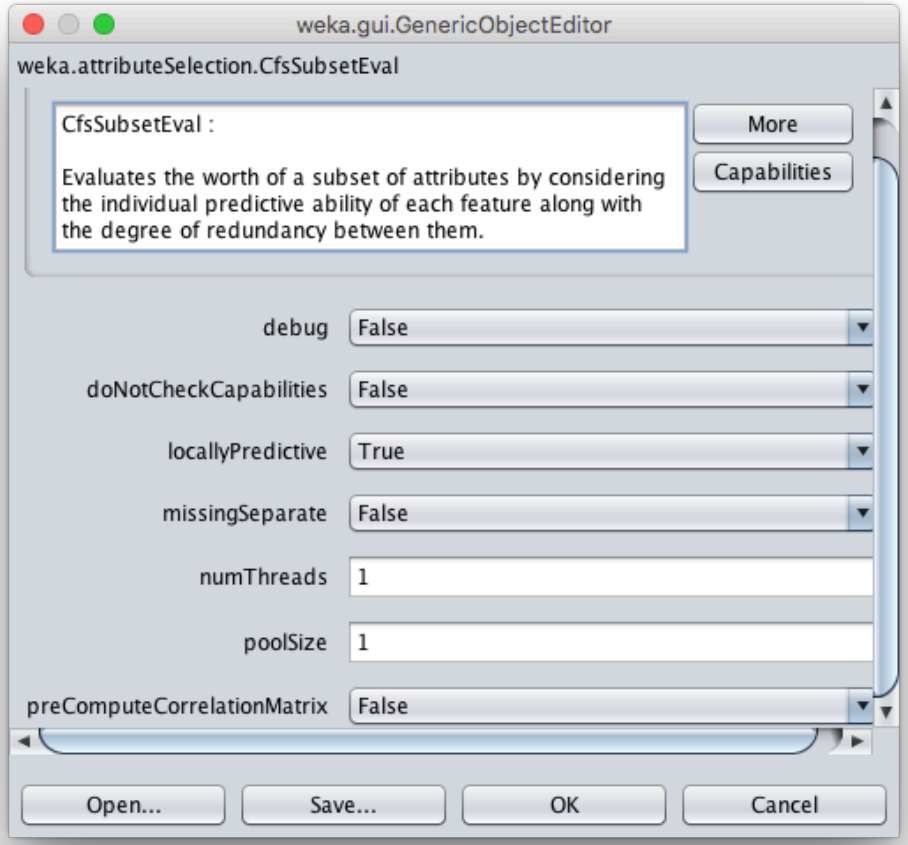



**Pre-procesamiento**

Además de utilizar “WordToStringVector” se agregó el filtro “Attribute Selection” el cuál evalúa por diferentes métodos tanto de evaluación como de búsqueda, sin embargo se utilizaron estos dos:

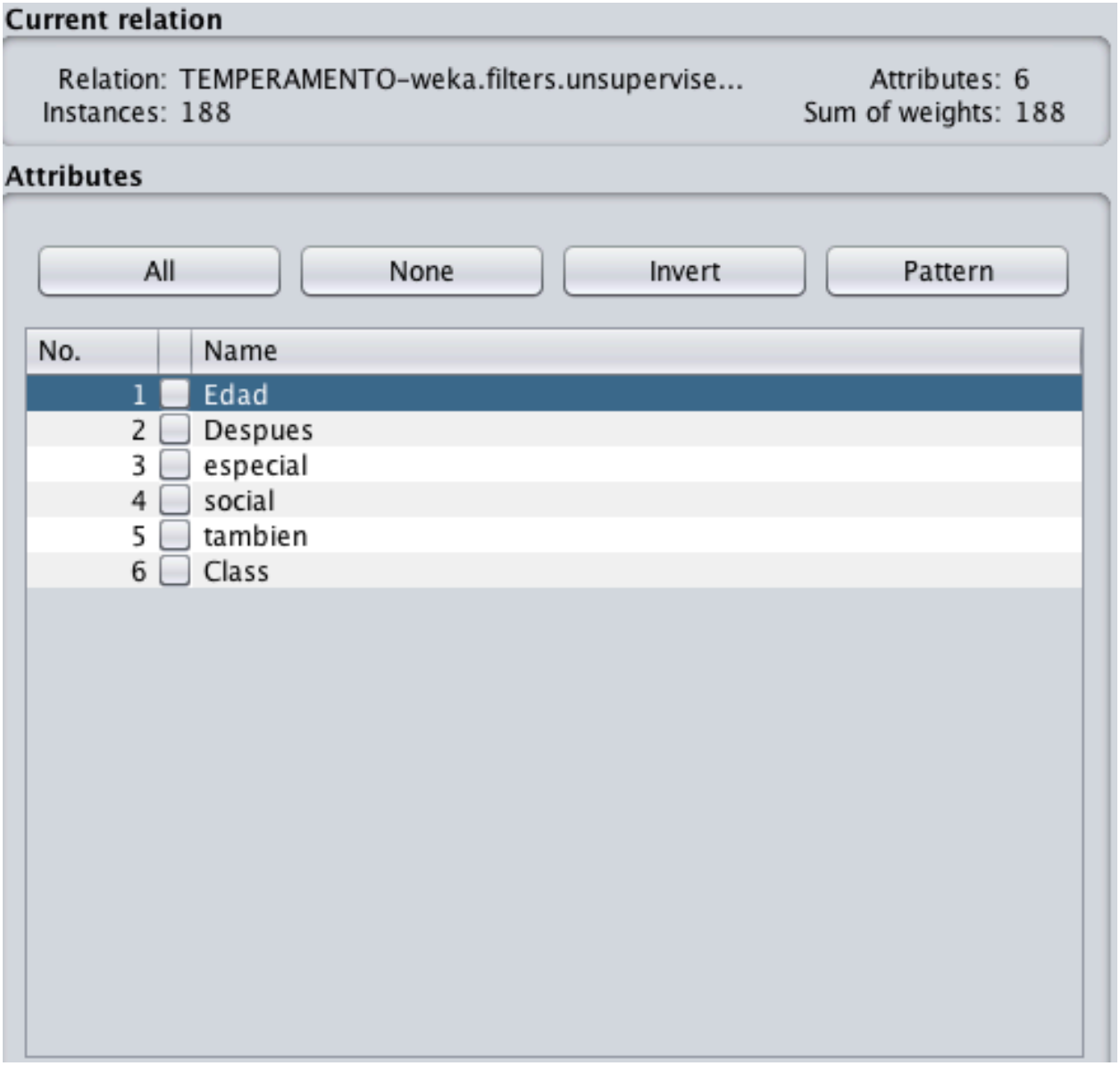



weka.gui.GenericObjectEditor
weka.attributeSelection.CfsSubsetEval
CfsSubsetEval :
Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
More
Capabilities
debug
False
doNotCheckCapabilities
False
locallyPredictive
True
missingSeparate
False
numThreads
1
poolSize
1
preComputeCorrelationMatrix
False
Open...
Save...
OK
Cancel

Este filtro se usó en el archivo equilibrado y el no equilibrado, quedando esto en el No equilibrado:

Y solo 6 atributos en el equilibrado.




Dado que en ambos casos salió que la Edad es un atributo importante se realizaron pruebas también sin este atributo, para corroborar que los resultados no estén siendo incorrectos o que se esté haciendo mal la clasificación.

## Clasificación

Al finalizar el pre-procesamiento se probó con la mayoría de métodos de clasificación en el archivo original y el archivo equilibrado con el atributo “Edad” y sin este mismo, sin embargo en la mayoría de los casos arrojó un porcentaje de respuestas correctas menores de 20% - 30% a 5 pliegos y mejoró un poco pero no pasando de 50% a 10 pliegos, a continuación se muestran las 3 mejores clasificaciones de los archivos sin equilibrar y equilibrados:

## CON EDAD

| BALANCEADO<br>47 T + 47 E + 47 S + 47 P | | | | | |
|---|---|---|---|---|---|
| **Clasificador** | **Precisión** | **Recuperación** | **Área bajo la Curva** | **Exactitud** | |
| | | | | **Correctos** | **Incorrectos** |
| **SIN ATTRIBUTE SELECTION (Atributos)** | | | | | |
| **BayesNet** | 0.319 | 0.330 | 0.581 | 62 32.98% | 126 67.02% |
| **JRip** | 0.246 | 0.250 | 0.502 | 47 25.00% | 141 75.00% |
| **Naive Bayes** | 0.235 | 0.234 | 0.468 | 44 23.40% | 144 76.60% |
| **CON ATTRIBUTE SELECTION (Atributos)** | | | | | |
| **Naive Bayes** | 0.433 | 0.394 | 0.610 | 74 39.36% | 114 60.64% |
| **Logistic** | 0.389 | 0.394 | 0.602 | 74 39.36% | 114 60.64% |
| **BayesNet** | 0.310 | 0.314 | 0.545 | 59 31.38% | 129 68.62% |

| DESBALANCEADO<br>54 T + 50 E + 50 S + 47 P | | | | | |
|---|---|---|---|---|---|
| **Clasificador** | **Precisión** | **Recuperación** | **Área bajo la Curva** | **Exactitud** | |
| | | | | **Correctos** | **Incorrectos** |
| **SIN ATTRIBUTE SELECTION (1382 Atributos)** | | | | | |
| **BayesNet** | 0.327 | 0.333 | 0.583 | 67 33.33% | 134 66.67% |
| **HoeffdingTree** | 0.291 | 0.279 | 0.527 | 56 27.86% | 145 72.14% |
| **J48** | 0.255 | 0.254 | 0.515 | 51 25.37% | 150 74.63% |
| **CON ATTRIBUTE SELECTION (43 Atributos)** | | | | | |
| **NaiveBayes** | 0.366 | 0.413 | 0.633 | 83 41.29% | 118 58.71% |
| **Logistic** | 0.371 | 0.378 | 0.627 | 76 37.81% | 125 62.19% |
| **MultilayerPerceptron** | 0.366 | 0.358 | 0.631 | 72 35.82% | 129 64.18% |

## SIN EDAD

| BALANCEADO<br>47 T + 47 E + 47 S + 47 P = 188 | | | | | |
|---|---|---|---|---|---|
| **Clasificador** | **Precisión** | **Recuperación** | **Área bajo la Curva** | **Exactitud** | |
| | | | | **Correctos** | **Incorrectos** |
| **SIN ATTRIBUTE SELECTION (Atributos)** | | | | | |
| **BayesNet** | ? | 0.330 | 0.508 | 62 32.98% | 126 67.02% |
| **JRip** | 0.222 | 0.234 | 0.460 | 44 23.40% | 144 76.60% |
| **Naive Bayes** | 0.279 | 0.277 | 0.478 | 52 27.66% | 136 72.34% |
| **CON ATTRIBUTE SELECTION (Atributos)** | | | | | |
| **Naive Bayes** | 0.366 | 0.309 | 0.603 | 58 30.85% | 130 69.15% |
| **Logistic** | 0.397 | 0.340 | 0.606 | 64 34.04% | 124 65.96% |
| **BayesNet** | ? | 0.330 | 0.508 | 62 32.98% | 126 67.02% |

| DESBALANCEADO<br>54 T + 50 E + 50 S + 47 P = 201 | | | | | |
|---|---|---|---|---|---|
| **Clasificador** | **Precisión** | **Recuperación** | **Área bajo la Curva** | **Exactitud** | |
| | | | | **Correctos** | **Incorrectos** |
| **SIN ATTRIBUTE SELECTION (1382 Atributos)** | | | | | |
| **BayesNet** | ? | 0.323 | 0.499 | 65 32.34% | 136 67.66% |
| **HoeffdingTree** | 0.293 | 0.279 | 0.526 | 56 27.86% | 145 72.14% |
| **J48** | 0.255 | 0.254 | 0.515 | 51 25.37% | 150 74.63% |
| **CON ATTRIBUTE SELECTION (43 Atributos)** | | | | | |
| **NaiveBayes** | 0.626 | 0.363 | 0.573 | 73 36.32% | 128 63.68% |
| **Logistic** | 0.645 | 0.368 | 0.581 | 74 36.82% | 127 63.18% |
| **MultilayerPerceptron** | 0.658 | 0.373 | 0.594 | 75 37.31% | 126 62.69% |

## Conclusión

El indice más alto fue el de NaiveBayes con 41.29% de aciertos lo que demuestra que es uno de los mejores métodos de clasificación por el momento a pesar de que no se tienen los mejores datos para clasificar este tipo de texto.

Se obtuvo un índice muy bajo de aciertos en la clasificación de texto, lo que indica que hay que buscar otros métodos para encontrar el tipo de apego que existe en alumnos y profesores de universidad, se puede recurrir a psicólogos o hacer una investigación más a fondo para encontrar atributos posibles para mejorar la clasificación de texto de igual forma se puede optimizar la busqueda implementando más métodos aparte de estos que se tienen posiblemente programándolos en otro lenguaje de programación.

Por el contrario que se creía al principio, el Género del autor de cada biografía no impactó mucho en la clasificación y se obtuvo que la Edad sí.

Como se puede observar, el atributo “Edad” sí contribuye a que los resultados mejoren aunque en ningún caso se pudo lograr un índice aceptable de clasificación.

**Referencias**

Cs.waikato.ac.nz. (2018). *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. [online] Available at: https://www.cs.waikato.ac.nz/~ml/weka/

Ocw.uc3m.es. [online] Available at: http://ocw.uc3m.es/ingenieria-informatica/herramientas-de-la-inteligencia-artificial/contenidos/transparencias/TutorialWeka.pdf.