BENEMÉRITA UNIVERSIDAD AUTONOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

SISTEMA DE PREDICCIÓN DE EDAD Y GÉNERO A PARTIR DE MENSAJES CORTOS.

TESIS PROFESIONAL

Para obtener el título de:

LIC. EN CIENCIAS DE LA COMPUTACIÓN.

Presenta:

Luis Ángel Sanjuán Palafox

Asesor:

Dr. David Eduardo Pinto Avendaño

Coasesor:

Dr. Darnes Vilariño Ayala

Puebla Puebla, Agosto 2013

RESUMEN

En este trabajo de tesis se investigaron y se implementaron métodos para el tratamiento automático de información, en particular, aplicable al problema de predicción de género y edad de autores de textos cortos.

Estonces, dado un mensaje textual corto, se diseño un algoritmo que permite determinar si el texto fue escrito por una persona del sexo masculino o del sexo femenino (género). Adicionalmente, el algoritmo permite predecir rangos de edad del autor. Estos rangos se establecen de la manera siguiente: 10s (de 13 a 17 años), 20s (de 23 a 27 años) y 30s (de lo 33 a los 37 años).

Para poder llevar a cabo esta tarea, se considero el uso de técnicas de aprendizaje automatico. Así, dado un conjunto de textos evaluados manualmente (training), se implementaron lo métodos de aprendizaje automático como son (Maquinas de Soporte Vectorial, Naive Bayes y Arboles de decisión C4.5 “J48” para generar los modelos de clasificación de nuevos textos, es decir, de los cuales se quiere predecir el género y edad del autor.

Se comparó el resultado obtenido de cada uno de los modelos, tomando en cuenta la precisión con la que las clases fueron correctamente identificadas, además del tiempo (en segundos) que les tomó en generar el modelo de clasificación.

Por último todos los resultados obtenidos a través de las pruebas realizadas son concentrados en dos tablas, una Tabla de Características indicando el porcentaje individual y colectivo que esta aporta para la correcta clasificación así como también una Tabla de porcentaje de efectividad de cada modelo implementado. Mostrando el resultado y la elección del mejor para cada caso.

Tabla de contenido

CAPITULO 1. INTRODUCCIÓN.

1.1. DESCRIPCIÓN DEL PROBLEMA.

En este trabajo, nos enfocamos en el problema "Predicción de género y edad en documentos de texto". Específicamente, nos centramos en las comunicaciones basadas en texto a través de Internet, como lo son las conversaciones publicadas en Redes Sociales. Es sumamente importante poder determinar con alto grado de confiabilidad el género y edad de las personas que envían mensajes a través de Internet. En particular, cuando estas personas se están comunicando con niños y/o adolecentes. Los jóvenes tienen un alto riesgo de encontrarse con posibles depredadores sexuales, por ejemplo. Así, un sistema que prediga tanto el género como la edad, sería de mucha utilidad en cualquiera de las redes sociales que actualmente son ampliamente utilizadas.

De manera adicional, las empresas tienen un gran interés en saber cuál es el género y la edad de las personas que emiten comentarios sobre sus productos en las redes sociales. El hecho de conocer estas dos características, permitiría tomar en cuenta dichos comentarios con la finalidad de ajustar los productos de la empresa. Comentarios positivos enfatizan las cualidades del producto, mientras que los negativos, muestran las debilidades de estos. Una manera eficiente de llegar al mercado es por supuesto, saber qué tipo de cliente opinan (género y edad).

Actualmente existen foros de competencia a nivel internacional que motivan la participación de equipos de investigación alrededor del mundo con el claro objetivo de esclarecer cuales son los principales retos al abordar la tarea de predicción de género y edad. El laboratorio de pruebas denominado PAN^2 (9th evaluation la bon uncovering plagiarism, authorship, and social software misuse) que se llevara a cabo en el marco de la conferencia CLEF 2013 es un claro ejemplo de interés creciente que existe en el desarrollo de técnicas apropiadas para la predicción de género y edad en redes sociales.

1.2. ESTADO DEL ARTE.

En [1] Se analizan y se procesan las conversaciones obtenidas en salas de chat, observando variaciones en los campos, tales como tendencias, hábitos, actitudes, situaciones de culpa, y las intenciones del autor. Con el objetivo de evaluar la determinación del mismo. Para ello se utilizan los clasificadores, Naive Bayes, K-vecino más cercano y Maquinas de soporte vectorial.

En [2] Se presenta un sistema de identificación de género en conversaciones por chat en Turquía. El sistema adquiere los datos y automáticamente los compara con entidades conocidas para determinar el género del autor de la conversación. Para ello se realiza una función de discriminación simple y además utiliza la implementación de un analizador semántico para mejorar el porcentaje de aciertos.

En [3] Se estudian las características de los mensajes de chat y propone un enfoque indicativo basado en la clasificación para la detección del tema chat. A partir del estudio de las características de mensajes de chat proponiendo un enfoque usando diferentes técnicas como sensibilización de mensajes de chat y la extracción de características de textos. Naive Bayes, clasificación asociativa y Maquinas de soporte vectorial se emplean como clasificadores para identificar los temas de las sesiones de chat.

En [4] Se presenta una investigación sobre la identificación de género y lenguaje en correos electrónicos utilizando minería y atribución, tomando las características como son marcadores de estilo, características estructurales, preferencias por género, funciones del lenguaje y también Maquinas de soporte vectorial como algoritmos de aprendizaje. Para los experimentos realizados se utilizó un corpus que contiene miles de correos electrónicos de ambos géneros para las pruebas que se realizaron.

En [5] Se presenta una mejora para la clasificación de documentos implementando el algoritmo de k-vecinos más cercanos que es un algoritmo de aprendizaje basado en instancias, la mejora se basa en limitar el número de instancias a comparar, para así disminuir el cálculo del vecino más cercano ajustando los pesos aprendiendo a distinguir diferentes características ya detectadas en un conjunto de documentos ya clasificados. Los pesos se ajustan utilizando un algoritmo iterativo.

En [6] Se propone una metodología que puede determinar automáticamente el género de los usuarios del chat. Para clasificar utilizan K-vecinos más cercanos, arboles de decisión, redes neuronales y

máquinas de soporte vectorial. Recopilan información de ambos géneros de diferentes redes sociales para después entrenar con una secuencia para definirlos y por ultimo hacer pruebas para la predicción de género.

En [7] se describe el desarrollo de una herramienta mediante técnicas de los campos sicométricos y de aprendizaje automático, debido a que los correos tienen macro-estructurales que son características las cuales pueden ser medidas. Estas características se pueden utilizar junto con el algoritmo de aprendizaje de máquinas de soporte vectorial para clasificar o atribuir la autoría de los mensajes de correo electrónico a un autor proporcionando suficientes muestras de mensajes para la comparación.

En [8] Se evalúan tres técnicas aplicadas al problema federalista. Las técnicas examinadas son una aproximación multivariada a la riqueza de vocabulario, análisis de las frecuencias de aparición de conjuntos en común, palabras de alta frecuencia y el uso de un paquete de aprendizaje automático basado en un "algoritmo genético" para buscar expresiones relacionales que caracterizan estilos del autor.

En [9] Se investiga la factibilidad de predecir el género del autor de un documento de texto, utilizando la evidencia lingüística. Para este propósito, las técnicas de clasificación término y el estilo de la base se evalúan sobre una gran colección de mensajes de chat.

En [10] Se muestra que las técnicas de categorización automática de texto pueden explotar combinaciones de simples rasgos léxicos y sintácticos para inferir el género del autor de un documento formal. Las mismas técnicas se pueden utilizar para determinar si un documento es ficción o no ficción.

En [11] Se investigó el problema de predecir el género del autor de un documento de texto. En particular, centrado en las comunicaciones basadas en texto a través de Internet. Primero formulando el problema como un problema de clasificación de texto, en el que las palabras de un documento se utiliza para atribuir un género al autor del documento. En segundo lugar, investigando el efecto de rasgos estilísticos (por ejemplo, longitudes de palabra, el uso de signos de puntuación, y emoticones) para predecir el género.

1.3. OBJETIVOS.

Objetivo general:

Evaluar e implementar diferentes técnicas para la predicción automática del género y edad de un autor, dado un texto corto en español.

Objetivos específicos:

- Identificar las técnicas que reportan mejores resultados en la predicción de la edad (10s, 20s, 30s) de un autor.
- Identificar las técnicas que reportan mejores resultados en la predicción del género (masculino o femenino) de un autor.
- Estudiar diversos métodos de clasificación automática con la finalidad de identificar el que mejor se comporte para el problema en cuestión.
- Implementar un algoritmo para la predicción de género y edad.
- Evaluar el algoritmo implementado usando datos supervisados.

1.4. PREGUNTAS DE INVESTIGACIÓN

¿Es posible detectar características lingüísticas en textos, que nos puedan ayudar a predecir el género de su autor?

¿Es posible detectar características lingüísticas en textos, que nos puedan ayudar a predecir la edad de su autor?

¿Qué clasificador nos aportara un mayor porcentaje de efectividad para la identificación de género?

¿Qué clasificador nos aportara un mayor porcentaje de efectividad para la identificación de Edad?

1.5. ESTRUCTURA DEL DOCUMENTO.

El primer capítulo es una introducción sobre el planteamiento del problema y el estado del arte en el que se encuentra actualmente la predicción de edad y género en textos de Redes Sociales. Además de definir los objetivos de la tesis.

A continuación, en el segundo capítulo se estudian los tres algoritmos, sobre los que se trabajaron para clasificar una vez recopiladas y extraídas las características lingüísticas seleccionadas de nuestro corpus de prueba.

En el tercer capítulo nombramos a PAN (uncovering plagiarism, authorship, and social software misuse). Para continuar con la descripción de los datos evaluados, seguido por el conjunto de características seleccionadas y por último los resultados experimentales obtuvimos para la predicción de edad y género.

En el cuarto capítulo describimos la interpretación de los resultados obtenidos a manera de concluir con la investigación.

CAPITULO 2. MARCO TEÓRICO

2.1. CLASIFICADORES

[14] Para poder realizar el reconocimiento automático de los objetos se realiza una transformación que convierte un objeto del universo de trabajo en un vector [25] X cuyas N componentes se llaman características discriminantes o rasgos.

Estas características deben permitir discriminar a qué clases puede pertenecer cualquier objeto del universo de trabajo.

$$X = (x_1, x_2, \ldots x_N) \quad \text{con} \quad N \in \mathbf{N} \quad \text{y} \quad x_i \in \Re \quad \forall i = 1 \ldots N$$

El valor del vector de características para un objeto concreto se conoce como patrón. Es decir, un patrón es una instancia particular de un vector de características determinado. La determinación de las N características discriminantes es un proceso difícil que suele requerir del uso de la imaginación. En general, suelen usarse características como los momentos de los objetos a reconocer, alguna transformación de los mismos (Fourier, cosenos...), las propias imágenes, o cualquier característica que se pueda obtener de los objetos mediante algún procedimiento algorítmico. Una vez determinadas las características discriminantes para un problema concreto, la clasificación de un objeto comienza por la obtención de su patrón. El siguiente paso consiste en determinar la proximidad o grado de pertenencia de este patrón a cada una de las clases existentes. A este efecto se definen las Funciones discriminantes o funciones de decisión como aquellas funciones que asignan a un patrón un grado de semejanza respecto a cada una de las diferentes clases.

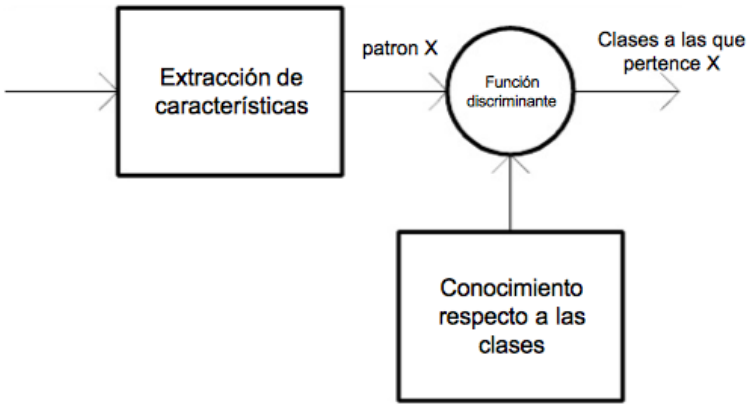


Figura 1. Esquema general del funcionamiento de un clasificador.

Clasificadores supervisados y no supervisados

Atendiendo a la información que se proporciona en el proceso de construcción del clasificador se puede hablar de dos tipos de clasificadores: con *maestro* o *supervisados*, sin maestros o *no supervisados*.

En los supervisados, la muestra la divide el maestro en las diferentes clases ya conocidas en las que se desea clasificar. A grandes rasgos las etapas en la construcción de un clasificador con maestro son: determinación de las clases, elección y test de las características discriminantes, selección de la muestra, cálculo de funciones discriminantes y test del clasificador.

En los no supervisados este proceso se realiza de manera automática, sin la necesidad de ningún supervisor externo. Para ello se emplean técnicas de agrupamiento, gracias a las cuales el sistema selecciona y aprende los patrones que poseen características similares, determinándose automáticamente las clases.

2.1.1. ÁRBOL DE DECISIÓN J48 (ALGORITMO C4.5).

[15] El algoritmo C4.5 fue desarrollado por JR Quinlan en 1993, como una extensión (mejora) del algoritmo ID3 que desarrollo en 1986.
En Weka se incluye una versión en Java de C4.5 llamado J48.

El algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye mediante la estrategia de profundidad-primero (depth-first).El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos. En cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos.

Los tres tipos de pruebas posibles propuestas por el C4.5 son:
La prueba "estándar" para las variables discretas, con un resultado y una rama para cada valor posible de la variable.

Una prueba más compleja, basada en una variable discreta, en donde los valores posibles son asignados a un número variable de grupos con un resultado posible para cada grupo, en lugar de para cada valor.

Si una variable A tiene valores numéricos continuos, se realiza una prueba binaria con resultados A <= Z y A > Z, para lo cual debe determinarse el valor límite Z.
Todas estas pruebas se evalúan de la misma manera, mirando el resultado de la proporción de ganancia, o alternativamente, el de la ganancia resultante de la división que producen. Ha sido útil agregar una restricción adicional: para cualquier división, al menos dos de los subconjuntos Ci deben contener un número razonable de casos. Esta restricción, que evita las subdivisiones casi triviales, es tenida en cuenta solamente cuando el conjunto C es pequeño.

CARACTERÍSTICAS DEL ALGORITMO C4.5

• Permite trabajar con valores continuos para los atributos, separando los posibles resultados en 2 ramas $A_i<=N$ y $A_i>N$.
• Los árboles son menos frondosos, ya que cada hoja cubre una distribución de clases no una clase en particular.
• Utiliza el método "divide y vencerás" para generar el árbol de decisión inicial a partir de un conjunto de datos de entrenamiento.
• Se basa en la utilización del criterio de proporción de ganancia (gain ratio), definido como $I(X_i,C)/H(X_i)$. De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección.
• Es Recursivo.

HEURÍSTICA

Utiliza una técnica conocida como Gain Ratio (proporción de ganancia). Es una medida basada en información que considera diferentes números (y diferentes probabilidades) de los resultados de las pruebas.

ATRIBUTOS

Atributos de valores continuos: Inicialmente el algoritmo ID3 se planteó para atributos que presentaban un número discreto de valores. Podemos fácilmente incorporar atributos con valores continuos, simplemente dividiendo estos valores en intervalos discretos, de forma que el atributo tendrá siempre valores comprendidos en uno de estos intervalos.

Medidas alternativas en la selección de atributos: Al utilizar la ganancia de información estamos introduciendo involuntariamente un sesgo que favorece a los atributos con muchos valores distintos. Debido a que dividen el conjunto de ejemplos en muchos subconjuntos, la ganancia de información es forzosamente alta. Sin embargo, estos atributos no son buenos predictores de la función objetivo para nuevos ejemplos. Una medida alternativa que se ha usado con éxito es la "gain ratio".

Atributos con valores perdidos: En ciertos casos existen atributos de los cuales conocemos su valor para algunos ejemplos, y para otros no. Por ejemplo una base de datos médica en la que no a todos los pacientes se les ha practicado un análisis de sangre. En estos casos lo más común es estimar el valor basándose en otros ejemplos de los que sí conocemos el valor. Normalmente se fija la atención en los demás ejemplos de ese mismo nodo. Así, al ejemplo de valor desconocido se le da el valor que más aparezca en los demás ejemplos.

Atributos con pesos diferentes: En algunas tareas de aprendizaje los atributos pueden tener costes asociados. Por ejemplo, en una aplicación médica para diagnosticar enfermedades podemos tener atributos como temperatura, resultado de la biopsia, pulso, análisis de sangre, etc., que varían significativamente en su coste, monetario y relativo a molestias para el paciente.
Ventajas respecto al algoritmo ID3

MEJORAS DEL ALGORITMO C4.5

- Evitar Sobreajuste de los datos.
- Determinar qué tan profundo debe crecer el árbol de decisión.
- Reducir errores en la poda.
- Condicionar la Post-Poda.
- Manejar atributos continuos.
- Escoger un rango de medida apropiado.
- Manejo de datos de entrenamiento con valores faltantes.
- Manejar atributos con diferentes valores.
- Mejorar la eficiencia computacional.

SOBREAJUSTE (OVERFITTING) A medida que se añaden niveles AD, las hipótesis se refinan tanto que describen muy bien los ejemplos utilizados en el aprendizaje, pero el error de clasificación puede aumentar al evaluar los ejemplos. Es decir, clasifica muy bien los datos de entrenamiento pero luego no sabe generalizar al conjunto de prueba. Es debido a que aprende hasta el ruido del conjunto de entrenamiento, adaptándose a las regularidades del conjunto de entrenamiento.

Este efecto es, por supuesto, indeseado. Hay varias causas posibles para que esto ocurra. Las principales son:

- Exceso de ruido (lo que se traduce en nodos adicionales)
- Un conjunto de entrenamiento demasiado pequeño como para ser una muestra representativa de la verdadera función objetivo. Hay varias estrategias para evitar el sobreajuste en los datos. Pueden ser agrupadas en dos clases:

1. Estrategias que frenan el crecimiento del árbol antes de que llegue a clasificar perfectamente los ejemplos del conjunto de entrenamiento.

2. Estrategias que permiten que el árbol crezca completamente, y después realizan una poda. POST PRUNNING (POST PODA) Es una variante de la poda y es usada por el C4.5. Consiste en una vez generado el árbol completo, plantearse qué es lo que se debe "podar" para mejorar el rendimiento y de paso obtener un árbol más corto. Pero además el C4.5 convierte el árbol a un conjunto de reglas antes de podarlo. Hay tres razones principales para hacer esto:

- Ayuda a distinguir entre los diferentes contextos en los que se usa un nodo de decisión, debido a que cada camino de la raíz a una hoja se traduce en una regla distinta.
- Deja de existir la distinción entre nodos que están cerca de la raíz y los que están lejos. Así no hay problemas para reorganizar el árbol si se poda un nodo intermedio.
- Mejora la legibilidad. Las reglas suelen ser más fáciles de entender. ESTRUCTURAS UTILIZADAS EN EL ALGORITMO C4.5 El C4.5 forma parte de la familia de los TDIDT (Top Down Induction Trees), junto con antecesor el ID3. El C4.5 se basa en el ID3, por lo tanto, la estructura principal de ambos métodos es la misma. El C4.5 construye un árbol de decisión mediante el algoritmo "divide y vencerás" y evalúa la información en cada caso utilizando los criterios de Entropía, Ganancia o proporción de ganancia, según sea el caso. Por otra parte, los árboles de decisión pueden entenderse como una representación de los procesos involucrados en las tareas de clasificación. Están formados por:

Nodos: Nombres o identificadores de los atributos.
Ramas: Posibles valores del atributo asociado al nodo.
Hojas: Conjuntos ya clasificados de ejemplos y etiquetados con el nombre de una clase.

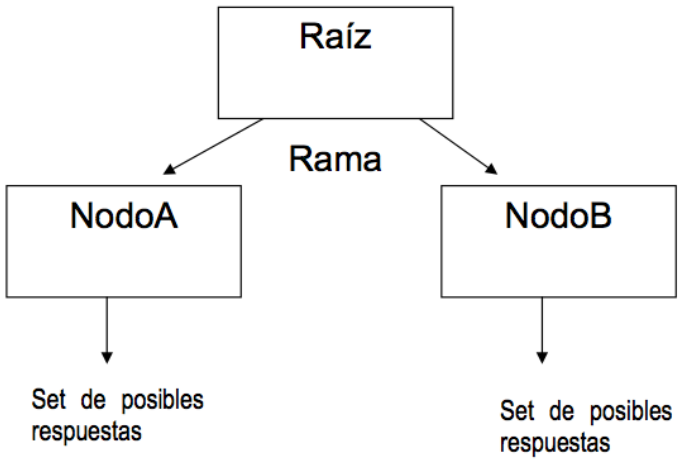
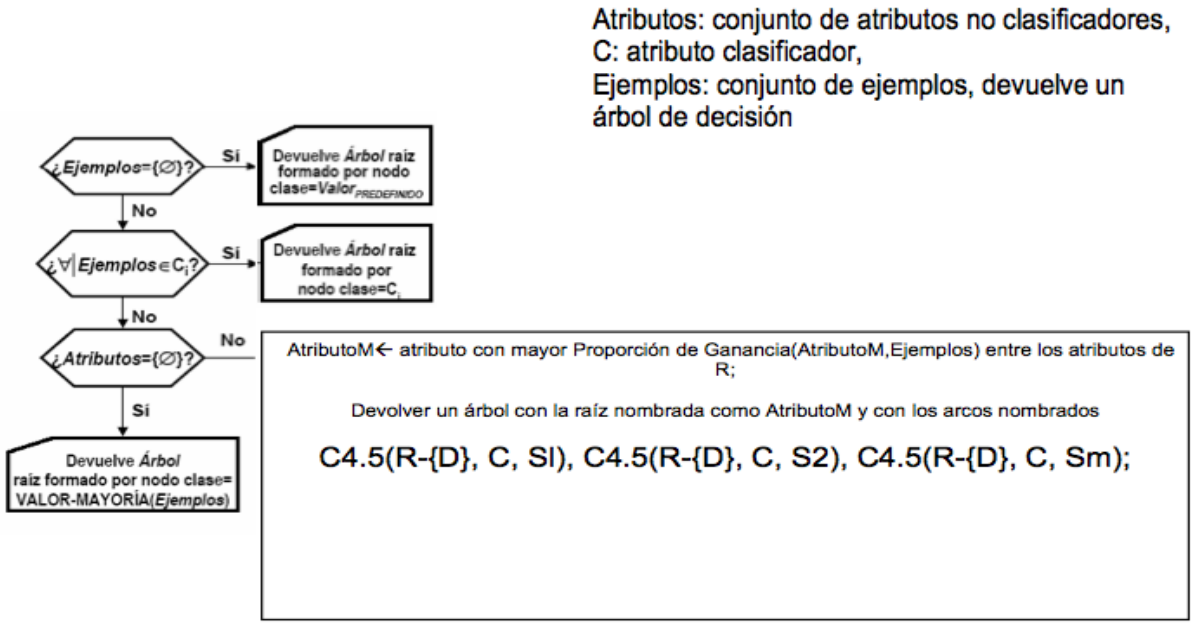


Figura 2. Ejemplo aplicado de Árbol de Decisión adaptado para C4.5

PSEUDOCODIGO DE C4.5

```
  Función C4.5

R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento, devuelve un árbol de decisión Comienzo

Si S está vacío,
      Devolver un único nodo con Valor Falla; ‘para formar el nodo raíz

Si todos los registros de S tienen el mismo valor para el atributo clasificador,
      Devolver un único nodo con dicho valor; ‘un único nodo para todos

Si R está vacío,
      Devolver un único nodo con el valor más frecuente del atributo
      Clasificador en los registros de S [Nota: habrá errores, es decir,
      Registros que no estarán bien clasificados en este caso];

Si R no está vacío,
      D <- atributo con mayor Proporción de Ganancia (D,S) entre los
      atributos de R;
      Sean {dj | j=1,2,...., m} los valores del atributo D;
      Sean {dj | j=1,2,...., m} los subconjuntos de S correspondientes a los
      valores de dj respectivamente;

      Devolver un árbol con la raíz nombrada como D y con los arcos
      nombrados d1, d2,....,dm, que van respectivamente a los árboles
      C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);

Fin
```

Diagrama genérico de algoritmo c4.5



ESTIMACIÓN DE LA PROPORCIÓN DE ERRORES PARA LOS ÁRBOLES DE DECISIÓN

Una vez podados, las hojas de los árboles de decisión generados por el C4.5 tendrán dos números asociados: N y E. N es la cantidad de casos de entrenamiento cubiertos por la hoja, y E es la cantidad de errores predichos si un conjunto de N nuevos casos fuera clasificados por el árbol. La suma de los errores predichos en las hojas, dividido el número de casos de entrenamiento, es un estimador inmediato del error de un árbol podado sobre nuevos casos. El C4.5 es una extensión del ID3 que acaba con muchas de sus limitaciones. Por ejemplo, permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas: una para aquellos A_i<=N y otra para A_i>N. Además, los árboles son menos frondosos porque cada hoja no cubre una clase en particular sino una distribución de clases, lo cual los hace menos profundos y menos frondosos. Este algoritmo fue propuesto por Quinlan en 1993. El C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de profundidad-primero (depth-first). Antes de cada partición de datos, el algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos.

2.1.2. NAIVE BAYES.

[16] El paradigma clasificatorio en el que se utiliza el teorema de Bayes en conjunción con la hipo´tesis de independencia condicional de las variables predictoras dada la clase se conoce bajo diversos nombres que incluyen los de idiota Bayes (Ohmann y col. 1988), Naive Bayes (Kononenko, 1990), simple Bayes (Gammerman y Thatcher, 1991) y Bayes independiente (Todd y Stamper, 1994).

A pesar de tener una larga tradición en la comunidad de reconocimiento de patrones (Duda y Hart, 1973) el clasificador Naive Bayes aparece por primera vez en la literatura del aprendizaje automático a finales de los ochenta (Cestnik y col. (1987)) con el objetivo de comparar su capacidad predictiva con la de métodos más sofisticados. De manera gradual los investigadores de esta comunidad de aprendizaje automático se han dado cuenta de su potencialidad y robustez en problemas de clasificación supervisada.

El paradigma Naive Bayes, debe su nombre a las hipótesis tan simplificadoras –independencia condicional de las variables predictoras dada la variable clase– sobre las que se construye dicho clasificador. Partirtiendo del paradigma clásico de diagnóstico para, una vez comprobado que necesita de la estimación de un número de parámetros ingente, ir simplificando paulatinamente las hipótesis sobre las que se construye hasta llegar al modelo Naive Bayes. Veremos a continuación un resultado teórico que nos servirá para entender mejor las características del clasificador Naive Bayes.

Del Paradigma Clásico de Diagnóstico al Clasificador Naive Bayes

Se comienza recordando el teorema de Bayes con una formulación de sucesos, para posteriormente formularlo en términos de variables aleatorias. Una vez visto el teorema de Bayes, se presenta el paradigma clásico de diagnóstico, viéndose la necesidad de ir simplificando las premisas sobre las que se construye en aras de obtener paradigmas que puedan ser de aplicación para la resolución de problemas reales. El contenido de este apartado resulta ser una adaptación del material que Díez y Nell (1998) dedican al mismo.

Teorema de (Bayes, 1764) 1 Sean A y B dos sucesos aleatorios cuyas probabilidades se denotan por p(A) y p(B) respectivamente, verificándose que p(B) > 0. Supongamos conocidas las probabilidades a priori de los sucesos A y B, es decir, p(A) y p(B), así como la probabilidad condicionada del suceso B dado el suceso A, es decir p(B|A). La probabilidad a posteriori del

suceso A conocido que se verifica el suceso B, es decir p(A|B), puede calcularse a partir de la siguiente fórmula:

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A)p(B|A)}{p(B)} = \frac{p(A)p(B|A)}{\sum_{A'} p(A')p(B|A')}$$

La formulación del teorema de Bayes puede efectuarse también para variables aleatorias, tanto unidimensionales como multidimensionales.

Comenzando por la formulación para dos variables aleatorias unidimensionales que denotamos por X e Y, tenemos que:

$$p(Y = y|X = x) = \frac{p(Y = y)p(X = x|Y = y)}{\sum_{y'} p(Y = y')p(X = x|Y = y')}$$

El teorema de Bayes también puede ser expresado por medio de una notación que usa el número de componentes de cada una de las variables multidimensionales anteriores X e Y, de la siguiente manera:

$$p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) = p(Y_1 = y_1, \ldots, Y_m = y_m|X_1 = x_1, \ldots, X_n = x_n)$$
$$= \frac{p(Y_1 = y_1, \ldots, Y_m = y_m)p(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y_1, \ldots, Y_m = y_m)}{\sum_{y'_1,\ldots,y'_m} p(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y'_1, \ldots, Y_m = y'_m)p(Y_1 = y'_1, \ldots, Y_m = y'_m)}$$

En el problema de clasificación supervisada reflejado en la siguiente Tabla, tenemos que Y = C es una variable unidimensional, mientras que X = (X_1, . . . , X_n) es una variable n-dimensional.

	X_1	...	X_n	Y
$(\boldsymbol{x}^{(1)}, y^{(1)})$	$x_1^{(1)}$	...	$x_n^{(1)}$	$y^{(1)}$
$(\boldsymbol{x}^{(2)}, y^{(2)})$	$x_1^{(2)}$	...	$x_n^{(2)}$	$y^{(2)}$
...		...		
$(\boldsymbol{x}^{(N)}, y^{(N)})$	$x_1^{(N)}$	...	$x_n^{(N)}$	$y^{(N)}$

Problema de clasificación supervisada.

Vamos a plantear la formulación clásica de un problema de diagnóstico utilizando una terminología habitual en medicina. Es evidente que la terminología puede trasladarse a otras ramas de la ciencia y de la técnica, en particular a la ingeniería. La terminología a usar incluye términos como:

- hallazgo, con el cual nos referimos a la determinación del valor de una variable predictora X_r. Así por ejemplo x_r (valor de la variable X_r) puede estar representando la existencia de vómitos en un determinado enfermo;
- evidencia, denota el conjunto de todos los hallazgos para un determinado individuo. Es

decir x = (x_1, . . . , x_n) puede estar denotando (si n = 4) que el individuo en cuestión es joven, hombre, presenta vómitos y además no tiene antecedentes familiares;

- diagnóstico, denota el valor que toman las m variables aleatorias Y_1,...,Y_m, cada una de las cuales se refiere a una enfermedad;
- probabilidad a priori del diagnóstico, p(y) o p($Y_1 = y_1$, . . . , $Y_m = y_m$), se refiere a la probabilidad de un diagnóstico concreto, cuando no se conoce nada acerca de los hallazgos, es decir, cuando se carece de evidencia;
- probabilidad a posteriori de un diagnóstico, p(y|x) o p($Y_1 = y_1$, . . . , $Y_m = y_m | X_1 = x_1$, . . . , $X_n = x_n$), es decir, la probabilidad de un diagnóstico concreto cuando se conocen n hallazgos (evidencia).

En el planteamiento clásico del diagnóstico (véase la siguiente Tabla) se supone que los m diagnósticos posibles son no excluyentes, es decir, pueden ocurrir a la vez, siendo cada uno de ellos dicotómico. Para fijar ideas en relación con el ámbito médico, podemos pensar que cada uno de los m posibles diagnósticos no excluyentes se relaciona con una enfermedad, pudiendo tomar dos valores: 0 (no existencia) y 1 (existencia). Por lo que se refiere a los n hallazgos o síntomas, se representarán por medio de las n variables aleatorias X_1,...,X_n y tambíen asumiremos que cada variable predictora es dicotómica, con valores 0 y 1. El valor 0 en la variable X_i indica la ausencia del i- ésimo hallazgo o síntoma mientras que el valor 1 indica la presencia del hallazgo o síntoma correspondiente.

	X_1	...	X_n	Y_1	...	Y_m
$(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)})$	$x_1^{(1)}$	...	$x_n^{(1)}$	$y_1^{(1)}$	...	$y_m^{(1)}$
$(\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)})$	$x_1^{(2)}$	...	$x_n^{(2)}$	$y_1^{(2)}$	...	$y_m^{(2)}$
...		...			...	
$(\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})$	$x_1^{(N)}$	...	$x_n^{(N)}$	$y_1^{(N)}$	...	$y_m^{(N)}$

Problema clásico de diagnóstico.

El problema del diagnóstico consiste en encontrar el diagnóstico más probable a posteriori, una vez conocido el valor de la evidencia. En notación matemática el diagnóstico óptimo, (y_1^*, . . . , y_m^*) será aquel que verifique:

$$(y_1^*, \ldots, y_m^*) = arg \max_{(y_1,\ldots,y_m)} p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$$

Aplicando el teorema de Bayes para calcular p($Y_1 = y_1$, . . . , $Y_m = y_m | X_1 = x_1$,...,$X_n = x_n$), obtenemos:

$$p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$$
$$= \frac{p(Y_1 = y_1, \ldots, Y_m = y_m) p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m)}{\sum_{y_1', \ldots, y_m'} p(Y_1 = y_1', \ldots, Y_m = y_m') p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1', \ldots, Y_m = y_m')}$$

Veamos a continuación el número de parámetros que se deben estimar para poder especificar el paradigma anterior y de esa forma obtener el valor de (y_1^*, . . . , y_m^*). Es importante tener en cuenta que la estimación de cada uno de los parámetros anteriores se deberá efectuar a partir del fichero de N casos, reflejado en la Tabla anterior.

Para estimar p($Y_1 = y_1$, . . . , $Y_m = y_m$), y teniendo en cuenta que cada variable Y_i es dicotómica,

necesitaremos un total de $2_m - 1$ parámetros. De igual forma, por cada una de las distribuciones de probabilidad condicionadas, $p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m)$, se necesitan estimar $2_n - 1$ parámetros. Al tener un total de 2_m de tales distribuciones de probabilidad condicionadas, debemos estimar $(2_n - 1)2_m$ parámetros. Es decir, que el número total de parámetros necesarios para determinar un modelo concreto del paradigma clásico de diagnóstico es: $2_m - 1 + 2_m(2_n - 1)$. Para hacernos una idea del número de parámetros a estimar podemos consultar la siguiente Tabla, en la cual vemos de manera aproximada el número de parámetros a estimar para distintos valores de m (número de enfermedades) y n (número de hallazgos).

Número de parámetros a estimar, en función de m (número de enfermedades) y n (número de síntomas), en el paradigma clásico de diagnóstico.

m	n		parámetros
3	10	$\simeq$	$8 \cdot 10^3$
5	20	$\simeq$	$33 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{17}$

Ante la imposibilidad de poder estimar el elevado número de parámetros que se necesitan en el paradigma clásico de diagnóstico, en lo que sigue se simplificarán las premisas sobre las que se ha construido dicho paradigma.

En primer lugar vamos a considerar que los diagnósticos son excluyentes, es decir, que dos diagnósticos no pueden darse al unísono. Esto trae como consecuencia que en lugar de considerar el diagnóstico como una variable aleatoria m–dimensional, este caso pueda verse como una única variable aleatoria unidimensional siguiendo una distribución polinomial con m valores posibles.

Vamos a denotar por $X_1, \ldots, X_n$ a las n variables predictorias. Supongamos que todas ellas sean binarias. Denotamos por C la variable de diagnóstico, que suponemos puede tomar m posibles valores. La búsqueda del diagnóstico más probable a posteriori, c^*, una vez conocidos los síntomas de un determinado paciente, $x = (x_1, \ldots, x_n)$, puede plantearse como la búsqueda del estado de la variable C con mayor probabilidad a posteriori. Es decir

$$c^* = arg\,\max_c p(C = c | X_1 = x_1, \ldots, X_n = x_n)$$

El cálculo de $p(C = c | X_1 = x_1, \ldots, X_n = x_n)$ puede llevarse a cabo utilizando el teorema de Bayes, y ya que el objetivo es calcular el estado de C, c^*, con mayor probabilidad a posteriori, no es

$$p(C = c | X_1 = x_1, \ldots, X_n = x_n) \propto p(C = c) p(X_1 = x_1, \ldots, X_n = x_n | C = c)$$

necesario calcular el denominador del teorema de Bayes. Es decir,

Por tanto, en el paradigma en el que los distintos diagnósticos son excluyentes, y considerando que el número de posibles diagnósticos es m, y que cada variable predictoria X_i es dicotómica, tenemos que el número de parámetros a estimar es $(m-1)+m(2_n - 1)$, de los cuales:

$m - 1$ se refiere a las probabilidades a priori de la variable C;

$m(2_n - 1)$ se relacionan con las probabilidades condicionadas de cada posible combinación de las variables predictoras dado cada posible valor de la variable C.

La Tabla 6.4 nos da una idea del número de parámetros a estimar para distintos valores de m y n.

Vemos de nuevo que el número de parámetros a estimar sigue siendo elevado, de ahí que necesitamos imponer suposiciones más restrictivas para que los paradigmas

m	n		parámetros
3	10	$\simeq$	$3 \cdot 10^3$
5	20	$\simeq$	$5 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{15}$

Número de parámetros a estimar, en función de m (número de enfermedades) y n (número de síntomas), en el paradigma clásico de diagnóstico con diagnósticos excluyentes.

puedan convertirse en modelos implementables.

Vamos finalmente a introducir el paradigma Naive Bayes: diagnósticos excluyentes y hallazgos condicionalmente independientes dado el diagnóstico. El paradigma Naive Bayes se basa en dos premisas establecidas sobre las variables predictoras (hallazgos, síntomas) y la variable a predecir (diagnóstico). Dichas premisas son:

los diagnósticos son excluyentes, es decir, la variable C a predecir toma uno de sus m posibles valores: $c_1, \ldots, c_m$;

los hallazgos son condicionalmente independientes dado el diagnóstico, es decir, que si uno conoce el valor de la variable diagnóstico, el conocimiento del valor de cualquiera de los hallazgos es irrelevante para el resto de los hallazgos. Esta condición se expresa matemáticamente por medio de la fórmula:

$$p(X_1 = x_1, \ldots, X_n = x_n | C = c) = \prod_{i=1}^{n} p(X_i = x_i | C = c) \qquad (1)$$

ya que por medio de la regla de la cadena se obtiene:

$$\begin{aligned} p(X_1 = x_1, \ldots, X_n = x_n | C = c) \;=\; & p(X_1 = x_1 \mid X_2 = x_2, \ldots, X_n = x_n, C = c) \\ & p(X_2 = x_2 \mid X_3 = x_3, \ldots, X_n = x_n, C = c) \\ & \ldots p(X_n = x_n \mid C = c) \end{aligned}$$

Por otra parte teniendo en cuenta la independencia condicional entre las variables predictoras dada la variable clase, se tiene que:

$$p(X_i = x_i \mid X_{i+1} = x_{i+1}, \ldots, X_n = x_n, C = c) = p(X_i = x_i \mid C = c)$$

para todo i = 1, . . . , n. De ahı́ que se verifique la ecuación 1.

Por tanto, en el paradigma Naive Bayes, la búsqueda del diagnóstico más probable, c^*, una vez conocidos los síntomas $(x_1, \ldots, x_n)$ de un determinado paciente, se reduce a:

$$c^* = arg\,\max_c p(C = c|X_1 = x_1, \ldots, X_n = x_n)$$

$$= arg\,\max_c p(C = c) \prod_{i=1}^{n} p(X_i = x_i|C = c)$$

Suponiendo que todas las variables predictoras son dicotómicas, el número de parámetros necesarios para especificar un modelo Naive Bayes resulta ser $(m-1)+m_n$, ya que

- se necesitan $(m - 1)$ para´metros para especificar la probabilidad a priori de la variable C;
- para cada variable predictora X_i se necesitan m parámetros para determinar las distribuciones de probabilidad condicionadas.

Con los nu´meros reflejados en la siguiente Tabla, nos podemos hacer una idea del número de parámetros necesarios en función del número de posibles diagnósticos y del nímero de síntomas necesarios para especificar el paradigma Naive Bayes.

m	n	parámetros
3	10	32
5	20	104
10	50	509

Número de parámetros a estimar en el paradigma Naive Bayes en función del número de diagnósticos posibles (m) y del número de síntomas (n).

En el caso de que las n variables predictoras $X_1, \ldots, X_n$ sean continuas, se tiene que el paradigma Naive Bayes se convierte en buscar el valor de la variable C, que denotamos por c^*, que maximiza la probabilidad a posteriori de la variable C, dada la evidencia expresada como una instanciación de las variables $X_1,\ldots,X_n$, esto es, $x = (x_1,\ldots,x_n)$.

Es decir, el paradigma na¨ıve Bayes con variables continuas trata de encontrar c^* verificando:

$$c^* = arg\,\max_c p(C = c|X_1 = x_1, \ldots, X_n = x_n)$$

$$= arg\,\max_c p(C = c) \prod_{i=1}^{n} f_{X_i|C=c}(x_i|c)$$

donde $fX_i|C=c(x_i|c)$ denota, para todo i = 1,...,n, la función de densidad de la variable X_i condicionada a que el valor del diagnóstico sea c.

Suele ser habitual utilizar una variable aleatoria normal (para cada valor de C) para modelar el comportamiento de la variable X_i. Es decir, para todo c, y para todo $i \in \{1,\ldots,n\}$, asumimos

$$f_{X_i|C=c}(x_i|c) \rightsquigarrow \mathcal{N}(x_i; \mu_i^c, (\sigma_i^c)^2)$$

En tal caso el paradigma Naive Bayes obtiene c^*, como:

$$c^* = arg\,\max_c p(C = c) \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_i^c} e^{\frac{-1}{2}\left(\frac{x_i - \mu_i^c}{\sigma_i^c}\right)^2} \right]$$

En este caso el nu´mero de parámetros a estimar es (m − 1) + 2nm:

- m − 1 en relacio´n con las probabilidades a priori p(C = c);
- 2nm en relacio´n con las funciones de densidad condicionadas.

Finalmente puede ocurrir que algunos de los hallazgos se recojan en variables discretas mientras que otros hallazgos sean continuos. En tal caso hablaremos del paradigma Naive Bayes con variables predictoras continuas y discretas.

Supongamos que de las n variables predictoras, n_1 de ellas, $X_1, \ldots, Xn_1$, sean discretas, mientras que el resto $n - n_1 = n_2$, $Y_1, \ldots, Yn_2$, sean continuas. En principio al aplicar directamente la fórmula del paradigma Naive Bayes correspondiente a esta situación se obtiene:

$$p(c|x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i|c) \prod_{j=1}^{n_2} f(y_j|c)$$

Esta expresión puede propiciar el conceder una mayor importancia a las variables continuas, ya que mientras que $p(x_i|c)$ verifica $0 \leq p(x_i|c) \leq 1$, puede ocurrir que $f(y_j|c) > 1$. Con objeto de evitar esta situación, proponemos la normalización de la aportación de las variables continuas, dividiendo cada uno de los factores correspondientes por el $\text{máx}_{y_j} f(y_j|c)$. Obtenemos por tanto:

$$p(c|x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i|c) \prod_{j=1}^{n_2} \frac{f(y_j|c)}{\text{máx}_{y_j} f(y_j|c)} \qquad (2)$$

En el caso en que las funciones de densidad de las variables continuas condicionadas a cada posible valor de la variable clase sigan distribuciones normales, es decir si $Y_j|C=c \rightarrow N(y_j;\mu^c_j,(\sigma_j^c)2)$, se tiene que

$$\frac{f(y_j|c)}{\text{máx}_{y_j} f(y_j|c)} = \frac{\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{y_j-\mu_j^c}{\sigma_j^c}\right)^2}}{\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{\mu_j^c-\mu_j^c}{\sigma_j^c}\right)^2}} = e^{-\frac{1}{2}\left(\frac{y_j-\mu_j^c}{\sigma_j^c}\right)^2}$$

y la fórmula 2 se expresa de la manera siguiente:

$$p(c|x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i|c) \prod_{j=1}^{n_2} e^{-\frac{1}{2}\left(\frac{y_j-\mu_j^c}{\sigma_j^c}\right)^2}$$

2.1.3. MAQUINA DE SOPORTE VECTORIAL.

[17] La teoría de las Máquinas de Soporte Vectorial (SVM por su nombre en inglés Support Vector Machines) es una nueva técnica de clasificación y ha tomado mucha atención en años recientes. La teoría de la SVM está basada en la idea de minimización de riego estructural (SRM). En muchas aplicaciones, las SVM han mostrado tener gran desempeño, más que las máquinas de aprendizaje tradicional como las redes neuronales y han sido introducidas como herramientas poderosas para resolver problemas de clasificación.

Una SVM primero mapea los puntos de entrada a un espacio de características de una dimensión mayor (i.e.: si los puntos de entrada están en R^2 entonces son mapeados por la SVM a R^3) y encuentra un hyperplano que los separe y maximice el margen m entre las clases en este espacio como se aprecia en la Figura 3.

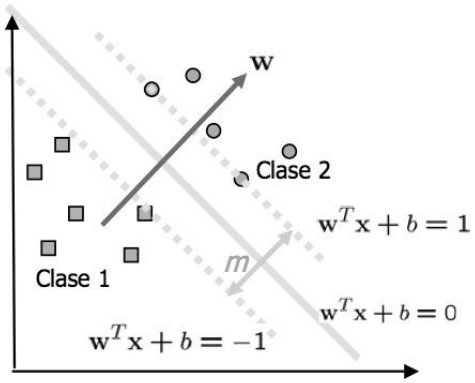


Figura 3. La frontera de decisión debe estar tan lejos de los datos de ambas clases como sea.

Maximizar el margen *m* es un problema de programación cuadrática (QP) y puede ser resuelto por su problema dual introduciendo multiplicadores de LaGrange. Sin ningún conocimiento del mapeo, la SVM encuentra el hyperplano óptimo utilizando el producto punto con funciones en el espacio de características que son llamadas *kernels.* La solución del hyperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados vectores de soporte.

Actualmente hay muchas aplicaciones que utilizan las técnicas de las SVM como por ejemplo las de OCR (Optical *Character Recognition)* por la facilidad de las SVMs de trabajar con imágenes como datos de entrada.

Caso linealmente separable

Supongamos que nos han dado un conjunto *S* de puntos etiquetados para entrenamiento como se aprecia en la Figura 4.

$$(y_1, x_1), \ldots, (y_l, x_l) \qquad (1)$$

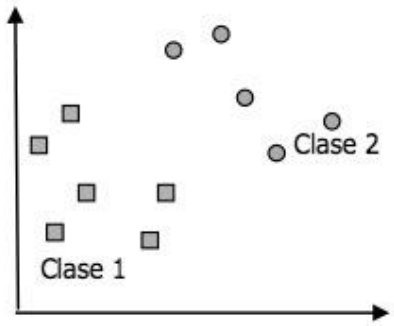


Figura 4. Caso linealmente separable.

Cada punto de entrenamiento $x_i \in R^N$ pertenece a alguna de las dos clases y se le da una etiqueta $y_i \in \{-1, 1\}$para i = 1, . . . , l. En la mayoría de los casos, la búsqueda de un hyperplano adecuado en un espacio de entrada es demasiado restrictivo para ser de uso práctico. Una solución a esta situación es mapear el espacio de entrada en un espacio de características de una dimensión mayor y buscar el hyperplano óptimo allí. Sea *z* =ϕ(x)la notación del correspondiente vector en el espacio de características con un mapeo ϕ de R^N a un espacio de características *Z.* Deseamos encontrar el hyperplano

$$w \cdot z + b = 0 \qquad (2)$$

Definido por el par (w, *b),* tal que podamos separar el cada punto de entrenamiento alguna de dos clases y se le ha dado una etiqueta punto x_i de acuerdo a la función

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1 & y_i = 1 \\ -1 & y_i = -1 \end{cases} \qquad (3)$$

Donde $w \in Z$ y $b \in R$. Más precisamente, el conjunto *S* se dice que es linealmente separable si existe (w, *b)* tal que las inecuaciones

$$\begin{cases} (w \cdot z_i + b) \geq 1, & y_i = 1 \\ (w \cdot z_i + b) \leq -1, & y_i = -1 \end{cases} \quad i = 1, \ldots, l \qquad (4)$$

Sean válidas para todos los elementos del conjunto *S.* Para el caso linealmente separable de *S,* podemos encontrar un único hyperplano óptimo, para el cual, el margen entre las proyecciones de los puntos de entrenamiento de dos diferentes clases es maximizado.

Caso no linealmente separable

Si el conjunto *S* no es linealmente separable, violaciones a la clasificación deben ser permitidas en la formulación de la SVM.

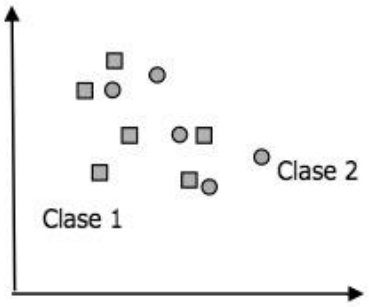


Figura 5. Caso no linealmente separable.

Para tratar con datos que no son linealmente separables, el análisis previo puede ser generalizado introduciendo algunas variables no-negativas ξi ≥ 0 de tal modo que (4) es modificado a

$$y_i(w \cdot z_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l. \qquad (5)$$

Los $\xi_i \neq 0$ en (5) son aquellos para los cuales el punto x_i no satisface (4). Entonces el término $\sum_{i=1}^{l}\xi_i$ puede ser tomado como algún tipo de medida del error en la clasificación.

El problema del hyperplano óptimo es entonces redefinido como la solución al problema

$$\begin{aligned} &\min \quad \left\{\frac{1}{2}w\cdot w + C\sum_{i=1}^{l}\xi_i\right\} \\ &s.a \quad y_i\left(w\cdot z_i + b\right) \geq 1-\xi_{i,} \quad i=1,\ldots,l \\ &\xi_i \geq 0, \quad i=1,\ldots,l \end{aligned} \qquad (6)$$

Donde *C* es una constante. El parámetro *C* puede ser definido como un parámetro de regularización. Este es el único parámetro libre de ser ajustado en la formulación de la SVM. El ajuste de éste parámetro puede hacer un balance entre la maximización del margen y la violación a la clasificación.

Buscando el hyperplano óptimo en (6) es un problema QP, que puede ser resuelto construyendo un Lagrangiano y transformándolo en el dual

$$\begin{aligned} &Max \quad W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j z_i\cdot z_j \\ &s.a \quad \sum_{i=1}^{l} y_i\alpha_i = 0, \quad 0\leq\alpha_i\leq C, \quad i=1,\ldots,l \end{aligned} \qquad (7)$$

Donde $\alpha = (\alpha_1,\ldots\alpha_l)$ es una vector de multiplicadores de LaGrange positivos asociados con las constantes en (5).

El teorema de Khun-Tucker juega un papel importante en la teoría de las SVM. De acuerdo a este teorema, la solución α_i del problema (7) satisface:

$$\overline{\alpha_i}\left(y_i\left(\overline{w}\cdot z_i + \overline{b}\right) - 1 + \overline{\xi_i}\right) = 0, \quad i=1,\ldots,l \qquad (8)$$

$$\left(C-\overline{\alpha_i}\right)\xi_i = 0, \quad i=1,\ldots l \qquad (9)$$

De esta igualdad se deduce que los únicos valores $\alpha_i \neq 0$ (9) son aquellos que para las constantes en (5) son satisfechas con el signo de igualdad. El punto xi correspondiente con $\alpha_i > 0$ es llamado vector de soporte. Pero hay dos tipos de vectores de soporte en un caso no separable. En el caso $0 < \alpha_i < C$, el correspondiente vector de soporte xi satisface las igualdades $y_i(w \cdot z_i + b) = 1$ y $\xi_i = 0$. En el caso $\alpha i = C$, el correspondiente ξ_i es diferente de cero y el correspondiente vector de soporte xi no satisface (4). Nos referimos a estos vectores de soporte como errores. El punto x correspondiente con $\alpha = 0$ es clasificado correctamente y está claramente alejado del margen de decisión. Figura 6.

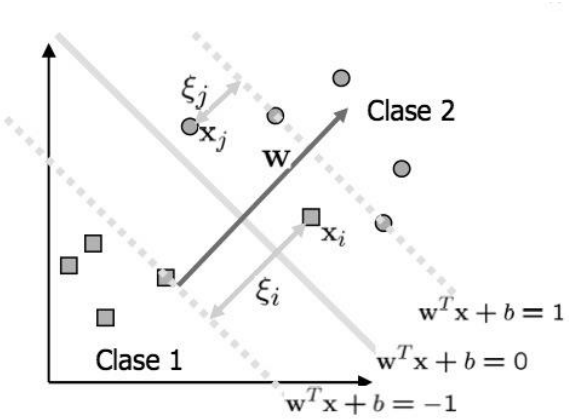


Figura 6. Aparición del parámetro de error ξ_i en el error de clasificación

Para construir el hyperplano óptimo $w \cdot z + b$, se utiliza

$$\overline{w} = \sum_{i=1}^{l} \overline{\alpha_i} y_i z_i \qquad (10)$$

Y el escalar *b* puede ser determinado de las condiciones de Kuhn-Tucker (9).

La función de decisión generalizada de (3) y (10) es tal que

$$f(x) = sign(w \cdot z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i z_i \cdot z + b\right) \quad (11)$$

2.2. SELECCCION DE CARACTERISTICAS

En esta fase se realiza la extracción de los datos relevantes para el análisis siguiendo los objetivos trazados a un principio. La calidad del conocimiento descubierto no sólo depende del algoritmo de clasificación utilizado, sino también de la calidad de los datos. Por lo tanto, después de la recopilación, el siguiente paso en el proceso es seleccionar y preparar el subconjunto de datos sobre los que se realizará la clasificación.

En general, en los procedimientos de selección de características se distinguen cuatro etapas esenciales:
Procedimiento de Selección: en esta etapa se determina el posible subconjunto de características para realizar la representación del problema

Función de Evaluación: en esta etapa se evalúa el subconjunto de características escogidas en el punto anterior.
Criterio de Detención: se chequea si el subconjunto seleccionado satisface el criterio de detención de la búsqueda.
Procedimiento de Validación: esta etapa se utiliza para verificar la calidad del subconjunto de características que se determinaron.
El procedimiento general de selección de características se ilustra en la Figura 5.
Los métodos de selección de características se clasifican desde el punto de vista de la manera en que se determina el nuevo subconjunto a evaluar, lo que conduce a 3 clases métodos.
Métodos Completos. Estos métodos examinan todas las posibles combinaciones de características. Son muy costosos computacionalmente (espacio de búsqueda de orden O(2N) para N características) pero se asegura encontrar el subconjunto óptimo de características.

Métodos Heurísticos. Utilizan una metodología de búsqueda de forma tal que no es necesario evaluar todos los subconjuntos de características. Ello significa una mayor velocidad del método, ya que el espacio de búsqueda es menor que en los métodos anteriores. Estos métodos no aseguran la obtención del mejor sub-conjunto.

Métodos Aleatorios. Son aquellos métodos que no tienen una forma específica de definir el subconjunto de características a analizar, sino que utilizan metodologías aleatorias. Con ello se produce una búsqueda probabilística en el espacio de características. El resultado obtenido utilizando este tipo de métodos dependerá del número de intentos, no asegurándose la obtención del óptimo.

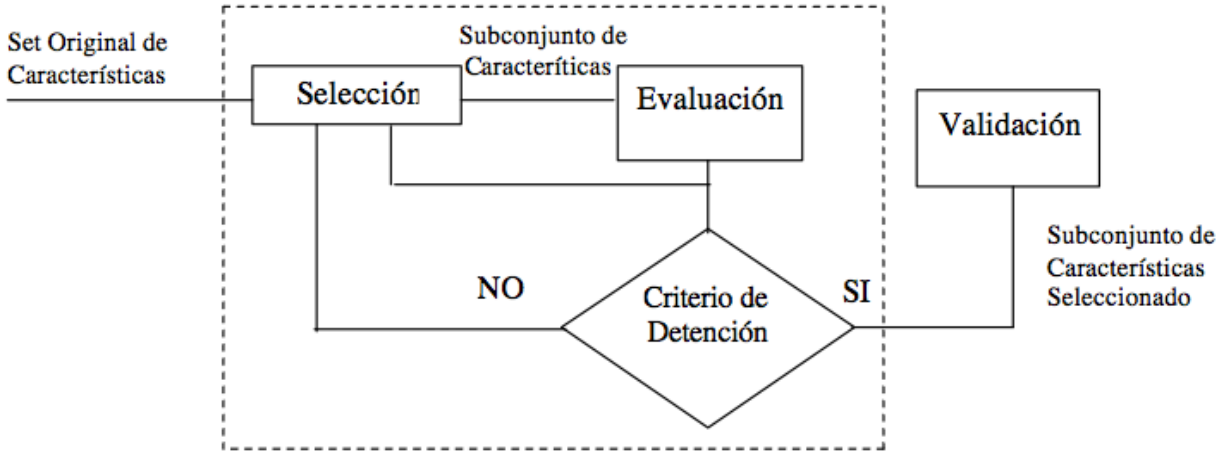


Figura 7. Procedimiento general de selección de características.

Desde el punto de vista de la función de evaluación, los procedimientos de selección de características se pueden clasificar en 2 categorías.

Métodos de filtraje. Estos son métodos donde el procedimiento de selección es realizado en forma independiente a la función de evaluación (clasificación). Se pueden distinguir 4 diferentes medidas: distancia, información, dependencia y consistencia.

Métodos dependientes (wrapped). En estos métodos el algoritmo de selección utiliza como medida la tasa de error del clasificador. Se obtienen generalmente mejores resultados que en el caso anterior, pero trae consigo un costo computacional mucho mayor.

2.3. PRE-PROCESAMIENTO DE DATOS

Cuando los datos se encuentran integrados lo primero que se debe realizar es un resumen de las características de atributos, con la ayuda de éstos resúmenes y características de los valores nominales se puede determinar fácilmente valores faltantes y valores erróneos.

Valores Faltantes:

Es importante detectar valores faltantes porque muchas tareas de minería de datos requieren datos completos para llevar a cabo un algoritmo. Para los valores faltantes se debe seguir las fases de detección y tratamiento. Tanto para la detección, como para su tratamiento posterior, es importante saber el porqué de los valores faltantes. Algunos valores faltantes expresan características relevantes, otros valores no existen o simplemente son datos incompletos.

Si se han conseguido establecer los datos faltantes e, idealmente, sus causas, se procederá a su tratamiento. Un método es reemplazar la información faltante por la media o la moda del atributo. Pero existen otras acciones que se mencionan a continuación:

- Ignorar, algunos algoritmos son robustos a datos faltantes
- Eliminar, filtrar o reemplazar toda la columna
- Filtrar la fila, claramente sesga los datos.
- Reemplazar el valor, se puede reemplazar por un valor que preserve la media o la varianza.
- Segmentar, se segmentan las tuplas por los valores que tienen disponibles.

Valores erróneos:

Son valores en la que una o más variables tienen valores que están significativamente fuera de la línea del valor promedio que es espera para esas variables. Del mismo modo que para los campos faltantes, se debe distinguir entre la detección y el tratamiento de los mismos, los tratamientos sobre datos erróneos son:

- Ignorar, algunos algoritmos son robustos a datos anómalos
- Filtrar la columna, solución extrema
- Filtrar la fila, puede sesgar los datos
- Reemplazar el valor, por el valor 'nulo' o predecir a partir de otros datos.
- Discretizar, transformar un valor continuo en uno discreto.

La transformación de datos engloba cualquier proceso que modifique la forma de los datos para que se refinen y ajusten a los requisitos de entrada del algoritmo de minería de datos. Las operaciones que

transforman los datos son: Reducción de dimensionalidad, aumento de dimensionalidad, discretización, numeración y normalización.

Discretización

La discretización o cuantización es la conversión de un valor numérico a un valor nominal ordenado. No obstante, el orden del atributo nominal puede ser preservado y utilizado por los pasos subsiguientes o bien puede olvidarse y tratarse el atributo como un valor nominal sin orden.

Pasos del pre-procesamiento.

El primer paso, fue tomar al azar un pequeño grupo de prueba para iniciar a trabajar (1,000 archivos) donde 500 son de mujeres y 500 de hombres de diferentes edades comprendidas entre 10 y 30 años de edad. Todo esto de un total de 20,000 archivos que componen el corpus de entrenamiento.

Los documentos XML contenidos en el corpus contienen una o varias conversaciones.

En la segunda etapa del pre-procesado de los datos consistió en extraer cada uno de los mensajes de cada interacción y resguardarla en un archivo de texto llamado “chat.txt”. Que se sobre escribe en cada iteración.

En la tercera etapa del pre-procesamiento se toquenizó nuestro conjunto de datos, para identificar posibles palabras, emoticones, Argots y signos de puntuación. También se normalizaron las palabras a minúsculas y se separaron los signos de puntuación de las palabras para una correcta identificación individual.

El cuarto paso fue toquenizar nuevamente cada una de las palabras, para extraer todos los tri-gramas que se pudieran generar dentro de cada texto. De tal forma que pudiéramos listar dichos resultados y compararlos entre grupos para así, poder identificar los más significativos.

El quinto paso consiste en agrupar los datos utilizando el perfil de los documentos, que incluye los datos del autor como el nombre y la edad para cada archivo. Esto con la única finalidad de poder hacer las observaciones correspondientes a cada grupo que se pretende clasificar. Para después realizar las pruebas convenientes en el corpus directamente sin ninguna agrupación.

EL sexto paso del pre-procesamiento fue la retroalimentación de los diccionarios que se mejoraban con cada una de las iteraciones ya que se reciclaba la información obtenida en cada una de

las pruebas, con el fin de incluir palabras, argots, caracteres y emoticones que no se habían incluido inicialmente.

El séptimo paso del pre-procesamiento, fue generar el conjunto de datos correspondientes a cada característica por grupo (Listados y Diccionarios), para finalmente comenzar a extraer las características del corpus completo comprendido por los 20,000 archivos que lo componen.

El octavo paso fue extraer todas las características de cada uno de los textos del corpus y concentrarlas en un único archivo ARFF el cual tiene la siguiente estructura:
Un archivo ARFF (Atributo-Relación File Format) es un archivo de texto ASCII que describe una lista de casos que comparten un conjunto de atributos.

```
% 1.  Título: Predicción de Edad y Género
% (A) Creador: Luis Ángel Sanjuán Palafox
% (B) Fecha: Junio 2013
%
@ RELACIÓN Conversaciones
@ ATRIBUTO FrecuenciaPalabrasCerradas {“Baja”, “Media”, “Alta”}
@ ATRIBUTO RiquezaVocabulario {“Pobre”, “Media”, “Alta”}
@ CLASE Genero{“Hombre”, “Mujer”}
@ DATA
   Baja, Media, Hombre
   Media, Pobre, Mujer
```

CAPITULO 3. AUTHOR PROFILING

[12] El Análisis de Autoría requiere de la clasificación de los textos en clases basadas en las opciones de estilo de sus autores. Más allá de la identificación de los autores y de las tareas de verificación de autor en el que se examinó el estilo de cada autor, el autor se distingue entre clases de perfiles de autores que estudian su aspecto sociológico, es decir, cómo el lenguaje es compartida por la gente. Esto ayuda en la identificación de aspectos de perfiles, tales como género, edad, lengua materna, o tipo de personalidad. Autor de perfiles es un problema de creciente importancia en aplicaciones en medicina forense, seguridad y marketing. Por ejemplo, desde la perspectiva de la lingüística forense a uno le gustaría ser capaz de conocer el perfil lingüístico del autor de un mensaje de texto de acoso (lenguaje utilizado por un determinado tipo de personas) e identificar ciertas características (lenguaje como prueba). Del mismo modo, desde el punto de vista de marketing, las empresas pueden estar interesados en saber, sobre la base del análisis de los blogs y reseñas de productos en línea, los datos demográficos de las personas que les gusta o no les gusta sus productos. La atención se centra en autor de perfiles en las redes sociales, ya que estamos sobre todo interesados en el lenguaje cotidiano y la forma en que refleja los procesos sociales y la personalidad básicas.

El reto: Teniendo en cuenta un documento, su tarea consiste en determinar la edad de su autor y su género.

3.1. DESCRIPCIÓN DETALLADA CLEF / PAN

[13] La Iniciativa de CLEF (Conferencia y laboratorios del Foro de Evaluación, antes conocido como Foro de Evaluación Cross-Language) es un organismo auto-organizado, cuya misión principal es promover la investigación, la innovación y el desarrollo de los sistemas de acceso a la información, con énfasis en multilingüe y multimodal información con diferentes niveles de la estructura. CLEF promueve la investigación y el desarrollo al proporcionar una infraestructura para:

- multilingüe y pruebas del sistema multimodal, ajuste y evaluación;
- investigación del uso y de los datos no estructurados, semi-estructurados, altamente estructurados, enriquecidos semánticamente en el acceso a la información;
- creación de colecciones de prueba reutilizables para la evaluación comparativa;
- exploración de nuevas metodologías de evaluación y formas innovadoras de utilización de los datos experimentales;

- discusión de los resultados, la comparación de los enfoques, el intercambio de ideas, y la transferencia de conocimiento.

La Iniciativa de CLEF se estructura en dos partes principales:

3. una serie de laboratorios de evaluación, es decir, los laboratorios para llevar a cabo la evaluación de los sistemas de acceso a la información y talleres para discutir y actividades piloto innovadores de evaluación;
4. una conferencia revisados en una amplia gama de temas, incluyendo

- investigación de la continuación de las actividades de los laboratorios de evaluación;
- experimentos con datos multilingües y multimodales y, en particular, pero no sólo, los datos resultantes de las actividades de CLEF;
- investigación en metodologías y desafíos de evaluación.

Desde 2000, el CLEF ha jugado un papel de liderazgo en la investigación y estimular la investigación en una amplia gama de áreas clave en el dominio de recuperación de información, llegando a ser bien conocido en la comunidad internacional IR. También ha promovido el estudio y la aplicación de metodologías de evaluación adecuadas para diversos tipos de tareas y los medios de comunicación. A través de los años, una amplia, fuerte y multidisciplinario comunidad científica se ha construido, que abarca y se extiende por las diferentes áreas de conocimiento necesarias para hacer frente a la expansión de las actividades de CLEF.

Los resultados fueron tradicionalmente se presentaron y debatieron en talleres anuales en conjunto con la Conferencia Europea de Bibliotecas Digitales (ECDL), que ahora se llama Teoría y práctica sobre bibliotecas digitales (TPDL).

Desde 2010, CLEF ha tomado la forma de un evento independiente, constituido por una conferencia revisada organizado conjuntamente con una serie de laboratorios de evaluación.

[12] PAN 2013 es la novena evaluación del laboratorio en descubrir el plagio, la autoría, y el mal uso de software social. PAN se llevará a cabo como parte de la conferencia de CLEF en Valencia, España, el 23-26 septiembre, 2013.

3.2. CONJUNTO DE DATOS PARA LA EVALUACIÓN.

[12] PAN puso a nuestra disposición un conjunto de datos de entrenamiento que consiste en documentos escritos en español. Con respecto a la edad, se consideró los mensajes en tres clases: 10s (13-17), 20s (23-27), y 30 (33-47).

Descripción de Corpus

El corpus se compone de documentos XML que contienen conversaciones (formato HTML) sobre muchos temas diferentes agrupadas por autor y etiquetada con su / su idioma, género y grupo de edad.

Idioma (Español), dos géneros (masculino y femenino), y tres grupos de edad (10s, 20s: 13-17: 23-27 y 30: 33-47).

Cada autor se presenta como un archivo XML independiente, el nombre de los cuales se proporciona información sobre el lenguaje, género y grupo de edad con el fin de facilitar las tareas de archivo, y agrupados por idioma en ES.

El nombre de cada documento XML tiene el formato:

UUID_lang_agegroup_gender.xml

Por ejemplo:

303232a213161ece822fe69176d48e58_en_20s_female.xml

Y cada archivo XML tiene el formato siguiente:

```
<autor lang="lang_code" gender="gender_code" age_group="age_group">
<conversations count="number_of_conversations_in_file">
id="UUID"> <conversation
[Contenido HTML original de la conversación]
</ Conversación>
id="UUID"> <conversation
[Contenido HTML original de la conversación]
</ Conversación>
      ....
</ Conversaciones>
</ Author>
```

3.3. DESCRIPCIÓN DE LAS CARACTERÍSTICAS USADAS.

Para la correcta clasificación de textos, hablando particularmente de textos cortos recopilados en Redes Sociales. Se identificaron diferentes tipos de características estilísticas y lingüísticas, la selección de dichas características puede afectar significativamente al rendimiento de los algoritmos de Clasificación.

Debido a que el lenguaje del corpus recuperado es el español, se requirió obtener un diccionario de la Lengua Española, en formato digital sin el significado o interpretación de las palabras. Agregando las conjugaciones de dichas palabras, además se incluyó palabras nativas del idioma hablado en nuestro país México.

La selección de las características estilísticas y lingüísticas, comenzó observando los textos de manera arbitraria, poniendo atención a posibles variaciones que pudieran marcar alguna diferencia, que nos ayudara a identificar la relación que existe entre dichas características y su autor.

Las características que se identificaron primero fueron: El uso y frecuencia de las palabras cerradas contenidas en el texto, las cuales se identifican por las vocales I - U que estas contienen y se llaman así debido a que al pronunciarlas, nuestra boca no se abre lo suficiente para pronunciar estas letras. A continuación seleccionamos como característica la ortografía y con ello la riqueza del vocabulario, identificando así las palabras bien escritas y pertenecientes a nuestro diccionario.
En cuanto a la riqueza, contabilizamos cada palabra diferente usada en el texto para después compararla con la diversidad de palabras contenidas en el diccionario.

Otra característica, fue el uso y frecuencia de los signos de puntuación los cuales, en algunos casos se abusaba del uso de los mismos, naturalmente obedecían a autores jóvenes que querían enfatizar de manera errónea palabras u oraciones en sus charlas en Redes Sociales.

Así mismo se eligió el uso de Argots de Internet, los cuales son una lista de términos y contracciones no estándares muy utilizados frecuentemente por los usuarios de Internet, particularmente en Redes Sociales. Incluimos también el uso de Emoticones los cuales son la unión de dos o más signos de puntuación, símbolos y caracteres que al observarlos de lado pueden formar gestos que usualmente los usuarios utilizan para notificar su estado de ánimo o representar objetos.

Se determinó medir la longitud de los textos como una característica que ayudara a definir a los autores.

Por otra parte las palabras contenidas en los textos del corpus, pueden descomponerse en (tri-gramas) los cuales dividen la palabra en tercias de letras, por ejemplo México (Méx-éxi-xic-ico). Esta es una propuesta muy relevante debido a que ayuda a identificar con un mayor porcentaje al autor.

Por ultimo haciendo una analogía con el entorno, muchas veces identificamos de manera automática al creador de un objeto por los errores y detalles que este le imprime a su obra de manera inconsciente. A lo que incluimos una lista de todos los errores ortográficos y aciertos más utilizados por cada grupo al que queremos clasificar.

3.3.1. PALABRAS CERRADAS.

Las palabras cerradas, se identifican por las vocales I - U que estas contienen y se llaman así debido a que al pronunciarlas, nuestra boca no se abre lo suficiente para pronunciar estas letras.

Ejemplos de palabras cerradas:

Asimismo. Debiendo. Dejando. Dicen. Diremos. Dar. Debiera. Dejara. Dices. Dirá. Da. Debieran. Dejaran. Diciendo. Dirán. Daba… etc.

A. USO DE PALABRAS CERRADAS.

En esta característica, identificamos dentro de cada texto el uso o no de las palabras cerradas, contabilizando una a una y al terminar de recorrer cada una de las palabras definimos por el contador si el autor uso o no uso palabras cerradas. Asignando así un identificador de los dos posibles “Usa”, “No-Usa”.

B. FRECUENCIA DE PALABRAS CERRADAS.

Una vez contabilizadas las palabras cerradas utilizadas, se evalúa con el promedio obtenido entre el mínimo uso y el máximo uso de estas, en todos los textos. Para posteriormente colocarle un identificador correspondiente de los siguientes; “Baja”, “Media” y “Alta”.

3.3.2. EVALUACIÓN DE ORTOGRAFÍA.

En esta característica se evalúa la correcta escritura de las palabras, las cuales se contabilizan a manera de buscar que cada palabra escrita en el texto, esté en el diccionario generado. El cual se compone por el diccionario de la Real Academia Española sin significados ni interpretaciones, más los conjugados de cada palabra revisados por un diccionario del editor de texto utilizado y además se incluyó palabras una lista de nombres propios.

Después de contabilizar el número de aciertos y el número total de palabras contenidas en el texto, se le da un identificador de acuerdo al porcentaje calculado entre el total de palabras y el número de aciertos ubicándolo con uno de estos identificadores: “Pobre”, “Promedio” o “Rico”.

3.3.3. USO DE SIGNOS DE PUNTUACIÓN.

Identificamos el número de signos de puntuación, contabilizando uno a uno conforme se van encontrando en el texto, si es que los hay, una vez leído por el sistema. Terminado este paso continuamos por evaluar, si el autor utilizo o no signos de puntuación, dándole uno de los identificadores; “Usa” o “No-Usa”.
Ejemplo de algunos signos de puntuación: { , . : ; ¿?¡! ... “” ‘ ‘ }

3.3.4. ARGOT DE INTERNET.

Debido al alto índice de deformación del lenguaje, particularmente por los jóvenes usuarios de la Internet que navegan en las Redes Sociales, quienes inventan frases, contracciones y palabras. Todo esto con el fin de ahorrar tiempo al escribir, han generado un listado de estas composiciones las cuales llamamos Argots de la Internet, vulgarmente la “Jerga de Internet”. La cual se aprovechó para tomarla como característica que ayudara a identificar a las personas jóvenes de las personas adultas. Por lo que contamos con un diccionario, en el cual incluimos un listado de Argots de Internet con la que comparamos palabra a palabra para contabilizar y posteriormente asignar un identificador; “No-Usa”, “Baja” o “Alta”.

Ejemplos de Argots de Internet:

ACG: (A call get) 'Recibí la llamada'. Usado para avisar que se recibió una llamada.

AEMJEFE: Kid nab de haloce.

AFAIK: (As Far As I Know) 'Hasta donde sé'. Usado para dar una información más bien escasa.

ASAP: (As Soon As Possible) 'Lo Mas pronto posible' en español, se usa cuando algo es prioritario. Ejemplo: "Cojan la bandera ASAP".

A4AF: (Asking for another file) 'Se le ha pedido otro archivo' en español, se usa en los programas P2P, cuando estas pidiendo un archivo y el usuario al que se le pide muestra A4AF en su estado. Ejemplo: eMule.

AFK: (Away from keyboard) En español es 'lejos del teclado', normalmente se usa en

MMORPGs, con jugadores que están ausentes.

ASL: (Age, Sex, Location) 'Edad, Sexo, Localización' Utilizada en los chats para dar a conocer estos datos o preguntárselos a alguna persona.

Admin: Apócope de Administrator/Administrador. Dícese de la persona que alguien elige para controlar un servidor, ya sea de un juego, un foro, etc. Esta persona posee el poder de *Banear, Kickear,* y hasta eliminar a los usuarios que estén en su servidor.

AFRW: "Away From Real World" lejos del mundo real, ya sea a personas que se imaginan que son parte y están dentro del juego.

AKA: "Also Known As", traducido "También conocido como".

APLH: "Apaga y Prende La Hueva". Apaga y enciende el computador/servidor después de una falla/error no recuperable.

ASDF: Se dice cuando no se tiene nada que decir pero se quiere tener la última palabra, o para crear spam en los foros de internet.

Banear: españolización del verbo inglés to ban (expulsar). Impedir el acceso a un usuario de un foro o un sitio por parte de la administración. Generalmente un usuario es *baneado* por incumplimiento reiterativo de las normas del lugar.

Brb: 'Be Right Back'; En traducción simple, 'vuelvo enseguida'.

BBL: 'Be back later'; En español, 'vuelvo más tarde'.

3.3.5. LONGITUD DEL MENSAJE.

Después de contar todas las palabras contenidas en el texto, evaluamos ese total de palabras con el promedio obtenido del máximo y mínimo de todos los textos, para después asignar uno de los posibles identificadores; “Corto”, “Medio” o “Largo” los cuales como su nombre lo dice, describen el tamaño de cada uno de los textos.

3.3.6. EMOTICONES.

Los Emoticones también conocidos como Smileys en inglés, son un conjunto de signos de puntuación, caracteres y símbolos. Los cuales al verlos de lado ya se izquierdo o derecho simulan algún gesto humano o en algunas ocasiones a algún objeto. Cada día se hacen más comunes en la Internet por lo que se diversifican con rapidez, a lo que muchos sitios como las Redes Sociales ya implementan identificadores a manera de intérpretes para cambiar dicho patrón de caracteres por imágenes más representativas.

Se generó un diccionario que contuviera a la gran mayoría de los emoticones registrados en la Internet, para así poder identificarlos uno a uno y contabilizarlos, a manera de poder asignar uno de los identificadores; “No-Usa”, “Baja” o “Alta. Con el cual nos ayudaría a definir a alguno de los grupos que estamos clasificando.

Icono	Significado
:-) :) :o) :] :3 :c) :> =] 8) =) :} :^) :つ)	Sonrisa
:-D :D 8-D 8D x-D xD X-D XD =-D =D =-3 =3 B^D	Risa
>:[:-(:(:-c :c :-< :つC :< :-[:[:{	Tristeza
:-|| :@	Enojo
:'-(:'(Q.Q	Llorando
:'-) :')	Lágrimas de felicidad
D:< D: D8 D; D= DX v.v D-':	Horror, asco, tristeza, gran consternación
>:O :-O :O °o° °O° :O o_O o_0 o.O 8-0	Sorpresa, shock, bostezo
:* :^* ('}{')	Beso, pareja besándose
;-) ;) *-) *) ;-] ;] ;D ;^) :-,	Guiño, sonrisa
>:P :-P :P X-P x-p xp XP :-p :p =p :-Þ :Þ :-b :b	Lengua fuera, atrevido / juguetón
>:\ >:/ :-/ :-. :/ :\ =/ =\ :L =L :S >.<	Escéptico, molesto, indeciso, inquieto y vacilante
:| :-|	Cara seria, sin expresión, indecisión
:$	Avergonzado, sonrojándose
:-X :X :-# :#	Labios sellados o usando aparatos ortopédicos
O:-) 0:-3 0:3 0:-) 0:) 0;^)	Ángel, santo, inocente
>:) >;) >:-)	Malvado
}:-) }:) 3:-) 3:)	Diabólico
o/\o ^5 >_>^ ^<_<	Chócala, choca esos cinco
|;-) |-O	Fresco, aburrido / bostezando

3.3.7. TRIGRAMAS.

Los trigramas son tercias de letras, por lo que descomponemos cada palabra en letras para formar dichos trigramas, los cuales generamos en listas por grupo, que después ordenamos por frecuencia de aparición eligiendo los más significativos y así formulamos nuestros diccionarios.

A. TRIGRAMAS POR (GÉNERO).

En este caso, primero dividimos el corpus en dos grupos, indicando el género para identificar los trigramas más significativos de cada uno y así poder formular dos diccionarios correspondientes a cada grupo. Para después comparar los trigramas resultantes de cada texto con los diccionarios generados, todo esto directamente en el corpus sin ninguna alteración. Al término de la evaluación de esta característica asignamos al resultado uno de los identificadores; “fem”, “mal” o “sin”. Con los que denotamos “fem” para indicar que hubo más incidencias en trigramas de género Femenino, “mal” para indicar que hubo más incidencias en trigramas de género Masculino y por ultimo “sin” que denota que no encontró ninguna coincidencia con alguno de los dos grupos.

B. TRIGRAMAS POR (EDAD).

Por otra parte, dividimos el corpus en tres partes, indicando por las tres diferentes edades a clasificar para identificar los trigramas más significativos de cada uno y así poder formular tres diccionarios correspondientes a cada grupo. Para después comparar los trigramas resultantes de cada texto con los diccionarios generados, todo esto directamente en el corpus sin ninguna alteración. Al término de la evaluación de esta característica, asignamos al resultado uno de los identificadores; “Bajo”, “Medio”, “Alto” o “Null”. Con los que denotamos a “Bajo” con el grupo de los 10 años, con “Medio” al grupo de los 20 años, con “Alto” al grupo de los 30 años y con “Null” si no encontraba al menos una coincidencia de los tres grupos anteriores.

3.3.8. ERRORES Y ACIERTOS.

Se recopilaron todas las faltas de ortografía, símbolos y caracteres divididos por grupos a clasificar, que después de calcular su frecuencia de incidencia, se elegían a los más significativos para así poder representar a cada uno de los grupos. Lo mismo se realizó con las palabras correctamente escritas, de igual forma divididas por grupo, calcular su frecuencia de incidencia para después elegir a las palabras más significativas que representarían a cada grupo.

A. ERRORES Y ACIERTOS (GÉNERO)

En este caso se dividió primero en dos grupos haciendo alusión al género correspondiente, se generaron las listas de los errores y palabras correctamente escritas que resultaron como más significativas y se procedió a identificar cada palabra contenida en este listado contra las obtenidas por los textos del corpus. Terminando de evaluar el número de incidencias se procedió a asignar un identificador; “F”, “M” o “S”. Con la “F” denotamos a que hubo mayor incidencia en el listado del grupo perteneciente al Femenino, con “M” denotamos que hubo mayor incidencia del grupo perteneciente al Masculino y con “S” indicamos que no tuvo al menos una coincidencia con alguno de los dos anteriores.

B. ERRORES Y ACIERTOS (EDAD).

En este caso se dividió en tres grupos haciendo alusión a los de 10 años, 20 años y 30 años correspondientemente, se generaron las listas de errores y palabras correctamente escritas que resultaron como más significativas y se procedió a identificar cada palabra contenida en este listado contra las obtenidas por los textos del corpus. Terminando de evaluar el número de incidencias se procedió a asignar un identificador; “E-10s”, “E-20s”, “E-30s” o “Sin”. Con lo que “E-10s” denotamos a los errores y aciertos del grupo de 10 años, con “E-20s” denotamos a los errores y aciertos del grupo de 20 años, con “E-30s” denotamos a los errores y aciertos del grupo de 30 años y por ultimo con “Sin” indicamos que no tuvo al menos una coincidencia con alguno de los tres anteriores.

C. ERRORES Y ACIERTOS (EDAD Y GÉNERO).

Para este procedimiento se dividió en seis grupos haciendo alusión a los de grupos de Edad y Género y sus posibles combinaciones: H-10 años, H-20 años, H-30 años, F-10 años, F-20 años y F-30 años correspondientemente, se generaron las listas de errores y palabras correctamente escritas que resultaron como más significativas y se procedió a identificar cada palabra contenida en este listado, contra las obtenidas por los textos del corpus.

Terminando de evaluar el número de incidencias se procedió a asignar un identificador; “H-10s”, “H-20s”, “H-30s”, “F-10s”, “F-20s”, “F-30s” o “Sin”. Con lo que “H-10s” denotamos a los errores y aciertos del grupo de hombres de 10 años, con “H-20s” denotamos a los errores y aciertos del grupo de hombres de 20 años, con “H-30s” denotamos a los errores y aciertos del grupo de hombres de 30 años, con lo que “F-10s” denotamos a los errores y aciertos del grupo de mujeres de 10 años, con “F-20s” denotamos a los errores y aciertos del grupo de mujeres de 20 años, con “F-30s” denotamos a los errores y aciertos del grupo de mujeres de 30 años y por ultimo con “Sin” indicamos que no tuvo al menos una coincidencia con alguno de los seis anteriores.

3.4. RESULTADOS EXPERIMENTALES

Con las características extraídas de cada una de las conversaciones del corpus parcial, se centralizaron en un solo archivo llamado Test01.arff, para después pasar a la etapa de Clasificación donde aplicamos tres tipos de clasificadores; Naive Bayes, Maquinas de Soporte Vectorial y J48. Estos algoritmos están ya implementados en el programa de distribución libre "WEKA" en su versión 3.6.8. El cual se usó para la generación del modelo de clasificación.

En un principio, solo se habían contemplado seis características; "palabrascerrdas", "correctas", "signospuntuacion", "argotinternet", "emoticones" y "riquezavocabulario". Las cuales se muestran en la siguiente estructura del primer archivo Test01.arff, que generamos después de extraer las características del corpus parcial de 1,000 elementos.

```
========================================
@relation Conversacion
@attribute palabrascerradas NUMERIC
@attribute correctas NUMERIC
@attribute signospuntuacion NUMERIC
@attribute argotinternet NUMERIC
@attribute emoticones NUMERIC
@attribute riquezavocavulario NUMERIC
@attribute clase {Male,Female}

@data
5, 6, 1, 0, 0, 0.00005,Male
11, 19, 12, 0, 0, 0.00025,Female
181, 287, 112, 0, 0, 0.00196,Male
17, 39, 38, 0, 0, 0.00043,Male
368, 592, 121, 0, 0, 0.00368,Male
10, 15, 2, 0, 0, 0.00015,Female
12, 12, 12, 0, 0, 0.00025,Female
91, 162, 132, 0, 0, 0.00140,Male
...
..
.
======================================
```

En esta primera prueba, utilizando el algoritmo J48 se obtuvo el siguiente porcentaje de efectividad:

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         587               58.5828 %
Incorrectly Classified Instances       415               41.4172 %
Kappa statistic                          0.1717
Mean absolute error                      0.4802
Root mean squared error                  0.4933
Relative absolute error                 96.046  %
Root relative squared error             98.6511 %
Total Number of Instances             1002

=== Detailed Accuracy By Class ===
```

```
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.653     0.481      0.576     0.653     0.612      0.574    Male
                   0.519     0.347      0.599     0.519     0.556      0.574    Female
Weighted Avg.      0.586     0.414      0.587     0.586     0.584      0.574

=== Confusion Matrix ===

   a   b   <-- classified as
 327 174 |   a = Male
 241 260 |   b = Female
```

Como se observa, se obtuvo un 58.58% de efectividad por lo nos dimos a la tarea de buscar y agregar nuevas características para mejorar, como lo fue la implementación de Trigramas. Con cada nueva mejora se realizaron nuevas pruebas, siempre conservando los archivos Test##.arff generados para comparar avances en el trabajo.

Cuando se implementó la característica “Trigramas” se generó el archivo Test05.arff, utilizando el algoritmo J48 nuevamente se obtuvo:

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         723               72.012  %
Incorrectly Classified Instances       281               27.988  %
Kappa statistic                          0.4398
Mean absolute error                      0.3883
Root mean squared error                  0.4415
Relative absolute error                 77.6612 %
Root relative squared error             88.2991 %
Total Number of Instances             1004

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.539     0.099      0.844     0.539     0.658      0.707    Male
                   0.901     0.461      0.662     0.901     0.763      0.707    Female
Weighted Avg.      0.72      0.281      0.753     0.72      0.711      0.707

=== Confusion Matrix ===

   a   b   <-- classified as
 270 231 |   a = Male
  50 453 |   b = Female
```

Se percibió un incremento en la efectividad ahora del 72.02%, después de obtener este resultado se observó por comparación con los Test anteriores, que además de la acertada integración de la característica “Trigramas”, la mejora de los listados y diccionarios aportaba un incremento

notable. Por lo que se procedió a retroalimentar los mismos, con recorridos superficiales a los archivos generados, que contenían los datos rescatados de cada una de las rutinas que se hacen para extraer las características. Realizando nuevamente pruebas con cada uno de los cambios, pero ahora ya implementados y probados en la totalidad del corpus.

Pero además, se buscó nuevamente otra característica que aportara en el incremento de efectividad, por lo que llegamos a la construcción del Test15.arff que mostro el siguiente resultado:

```
Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16264               81.32   %
Incorrectly Classified Instances      3736               18.68   %
Kappa statistic                          0.6264
Mean absolute error                      0.2478
Root mean squared error                  0.3549
Relative absolute error                 49.5518 %
Root relative squared error             70.9781 %
Total Number of Instances            20000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.815     0.188      0.812     0.815     0.813      0.886    Male
                 0.812     0.186      0.814     0.812     0.813      0.886    Female
Weighted Avg.    0.813     0.187      0.813     0.813     0.813      0.886

=== Confusion Matrix ===

    a    b   <-- classified as
 8145 1855 |    a = Male
 1881 8119 |    b = Female
```

Ya con un porcentaje de 81.32% se precedió a realizar las pruebas correspondientes con los demás clasificadores, buscando encontrar el que pudiera mostrar mejores ventajas, como son proyectar el mejor porcentaje de clasificación y en el menor tiempo en cuanto a la construcción del modelo de clasificación. Al mismo tiempo que se hacía para la predicción de edad y la combinación de estos (Edad / Género).

A continuación se describe con más detalle la implementación con cada uno de los tres algoritmos utilizados y su comparativa entre estos para cada caso.

3.4.1. GÉNERO

Como resultado de la implementación de la combinación de clases y la mejora que consistió en personalizar el listado de errores y aciertos, se generó el archivo Test16.arff con el cual se realizaron las pruebas siguientes:

Primero se comprobó nuevamente el porcentaje de efectividad utilizando el algoritmo J48 el cual se muestra a continuación:

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R15-16
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
              ErrorAciertosEdadGenero
              Genero
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
ErrorGenero = F: Female (6803.0/503.0)
ErrorGenero = M
|   ErrorAciertosEdadGenero = H-10s
|   |   ErrorEdad = E-10s: Male (369.0/42.0)
|   |   ErrorEdad = E-20s: Male (544.0/33.0)
|   |   ErrorEdad = E-30s
|   |   |   TriEdadGenero = TF10
|   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male (1.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Media: Male (5.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female (4.0/1.0)
|   |   |   TriEdadGenero = TF20: Female (2.0)
|   |   |   TriEdadGenero = TF30: Female (51.0/4.0)
|   |   |   TriEdadGenero = TM10: Male (50.0/17.0)
|   |   |   TriEdadGenero = TM20: Male (68.0/14.0)
|   |   |   TriEdadGenero = TM30: Male (41.0/2.0)
|   |   |   TriEdadGenero = Null: Male (49.0/13.0)
|   |   ErrorEdad = Sin: Male (16.0/3.0)
|   ErrorAciertosEdadGenero = H-20s: Male (658.0/11.0)
|   ErrorAciertosEdadGenero = H-30s: Male (1687.0/23.0)
|   ErrorAciertosEdadGenero = F-10s: Male (1276.0/94.0)
|   ErrorAciertosEdadGenero = F-20s
|   |   FrecuenciaPalabrasCerradas = Baja: Male (25.0/7.0)
|   |   FrecuenciaPalabrasCerradas = Media
|   |   |   LongitudMensaje = Corto: Male (0.0)
|   |   |   LongitudMensaje = Medio: Male (15.0/3.0)
```

```
|   |   |   LongitudMensaje = Largo
|   |   |   |   TriEdad = Bajo: Female (0.0)
|   |   |   |   TriEdad = Medio: Female (17.0/7.0)
|   |   |   |   TriEdad = Alto: Male (2.0)
|   |   |   |   TriEdad = Null: Female (0.0)
|   |   FrecuenciaPalabrasCerradas = Alta: Female (17.0/2.0)
|   ErrorAciertosEdadGenero = F-30s
|   |   ErrorEdad = E-10s: Female (1.0)
|   |   ErrorEdad = E-20s: Male (45.0/6.0)
|   |   ErrorEdad = E-30s
|   |   |   ArgotsInternet = No-Usa
|   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male (7.0/2.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Media: Female (12.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female (18.0/2.0)
|   |   |   ArgotsInternet = Baja: Female (13.0)
|   |   |   ArgotsInternet = Alta: Male (2.0)
|   |   ErrorEdad = Sin: Female (0.0)
|   ErrorAciertosEdadGenero = Sin: Male (2114.0/202.0)
ErrorGenero = S
|   TriEdadGenero = TF10: Female (339.0/100.0)
|   TriEdadGenero = TF20: Female (69.0/3.0)
|   TriEdadGenero = TF30: Female (1281.0/487.0)
|   TriEdadGenero = TM10
|   |   ErrorAciertosEdadGenero = H-10s
|   |   |   LongitudMensaje = Corto: Male (15.0/3.0)
|   |   |   LongitudMensaje = Medio: Male (9.0/3.0)
|   |   |   LongitudMensaje = Largo: Female (21.0/4.0)
|   |   ErrorAciertosEdadGenero = H-20s: Male (3.0)
|   |   ErrorAciertosEdadGenero = H-30s: Male (14.0/2.0)
|   |   ErrorAciertosEdadGenero = F-10s
|   |   |   ErrorEdad = E-10s
|   |   |   |   RiquezaVocabulario = Pobre: Male (2.0)
|   |   |   |   RiquezaVocabulario = Promedio: Female (4.0)
|   |   |   |   RiquezaVocabulario = Rico: Female (5.0)
|   |   |   ErrorEdad = E-20s: Male (27.0/9.0)
|   |   |   ErrorEdad = E-30s: Male (5.0/2.0)
|   |   |   ErrorEdad = Sin: Male (2.0)
|   |   ErrorAciertosEdadGenero = F-20s: Female (5.0)
|   |   ErrorAciertosEdadGenero = F-30s: Female (11.0/3.0)
|   |   ErrorAciertosEdadGenero = Sin
|   |   |   ArgotsInternet = No-Usa: Male (81.0/31.0)
|   |   |   ArgotsInternet = Baja
|   |   |   |   ErrorEdad = E-10s: Female (3.0)
|   |   |   |   ErrorEdad = E-20s: Male (9.0/3.0)
|   |   |   |   ErrorEdad = E-30s: Female (2.0)
|   |   |   |   ErrorEdad = Sin: Female (0.0)
|   |   |   ArgotsInternet = Alta: Female (1.0)
|   TriEdadGenero = TM20
|   |   ErrorAciertosEdadGenero = H-10s
|   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   ArgotsInternet = No-Usa: Male (28.0/6.0)
|   |   |   |   ArgotsInternet = Baja: Female (3.0/1.0)
|   |   |   |   ArgotsInternet = Alta: Male (0.0)
|   |   |   FrecuenciaPalabrasCerradas = Media: Male (9.0/2.0)
|   |   |   FrecuenciaPalabrasCerradas = Alta: Female (7.0/1.0)
|   |   ErrorAciertosEdadGenero = H-20s: Male (3.0)
|   |   ErrorAciertosEdadGenero = H-30s: Male (12.0/1.0)
|   |   ErrorAciertosEdadGenero = F-10s: Male (109.0/32.0)
|   |   ErrorAciertosEdadGenero = F-20s: Female (13.0/2.0)
|   |   ErrorAciertosEdadGenero = F-30s
|   |   |   ErrorEdad = E-10s: Female (0.0)
|   |   |   ErrorEdad = E-20s: Male (7.0/1.0)
|   |   |   ErrorEdad = E-30s: Female (5.0)
|   |   |   ErrorEdad = Sin: Female (0.0)
```

```
|   |   ErrorAciertosEdadGenero = Sin
|   |   |   ErrorEdad = E-10s: Male (7.0/2.0)
|   |   |   ErrorEdad = E-20s: Male (119.0/34.0)
|   |   |   ErrorEdad = E-30s: Female (15.0/4.0)
|   |   |   ErrorEdad = Sin: Male (33.0/10.0)
|   TriEdadGenero = TM30: Male (88.0/17.0)
|   TriEdadGenero = Null
|   |   ErrorAciertosEdadGenero = H-10s
|   |   |   ArgotsInternet = No-Usa
|   |   |   |   SignosPuntuacion = Usa
|   |   |   |   |   TriEdad = Bajo: Male (8.0/1.0)
|   |   |   |   |   TriEdad = Medio
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female (62.0/25.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   |   |   |   LongitudMensaje = Corto: Female (2.0)
|   |   |   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   |   |   ErrorEdad = E-10s: Female (7.0/3.0)
|   |   |   |   |   |   |   |   ErrorEdad = E-20s: Male (13.0/5.0)
|   |   |   |   |   |   |   |   ErrorEdad = E-30s: Male (5.0/1.0)
|   |   |   |   |   |   |   |   ErrorEdad = Sin: Female (3.0)
|   |   |   |   |   |   |   LongitudMensaje = Largo
|   |   |   |   |   |   |   |   TriGenero = fem: Female (2.0)
|   |   |   |   |   |   |   |   TriGenero = mal: Male (11.0/3.0)
|   |   |   |   |   |   |   |   TriGenero = sin: Female (2.0/1.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female (3.0/1.0)
|   |   |   |   |   TriEdad = Alto: Male (32.0/15.0)
|   |   |   |   |   TriEdad = Null: Female (30.0/12.0)
|   |   |   |   SignosPuntuacion = No-Usa: Male (38.0/13.0)
|   |   |   ArgotsInternet = Baja: Male (23.0/7.0)
|   |   |   ArgotsInternet = Alta: Male (0.0)
|   |   ErrorAciertosEdadGenero = H-20s: Male (18.0/4.0)
|   |   ErrorAciertosEdadGenero = H-30s: Male (92.0/17.0)
|   |   ErrorAciertosEdadGenero = F-10s
|   |   |   TriGenero = fem
|   |   |   |   ErrorEdad = E-10s: Female (37.0/11.0)
|   |   |   |   ErrorEdad = E-20s: Female (50.0/20.0)
|   |   |   |   ErrorEdad = E-30s: Female (3.0)
|   |   |   |   ErrorEdad = Sin
|   |   |   |   |   TriEdad = Bajo: Female (2.0)
|   |   |   |   |   TriEdad = Medio
|   |   |   |   |   |   RiquezaVocabulario = Pobre: Male (0.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio: Female (4.0/1.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico: Male (8.0)
|   |   |   |   |   TriEdad = Alto: Female (2.0)
|   |   |   |   |   TriEdad = Null: Male (3.0/1.0)
|   |   |   TriGenero = mal: Male (219.0/81.0)
|   |   |   TriGenero = sin
|   |   |   |   TriEdad = Bajo: Male (13.0/4.0)
|   |   |   |   TriEdad = Medio: Female (198.0/89.0)
|   |   |   |   TriEdad = Alto
|   |   |   |   |   ErrorEdad = E-10s: Female (8.0/3.0)
|   |   |   |   |   ErrorEdad = E-20s: Female (12.0/3.0)
|   |   |   |   |   ErrorEdad = E-30s: Male (3.0)
|   |   |   |   |   ErrorEdad = Sin: Male (10.0/1.0)
|   |   |   |   TriEdad = Null: Male (176.0/82.0)
|   |   ErrorAciertosEdadGenero = F-20s: Female (58.0/15.0)
|   |   ErrorAciertosEdadGenero = F-30s: Female (125.0/34.0)
|   |   ErrorAciertosEdadGenero = Sin
|   |   |   SignosPuntuacion = Usa
|   |   |   |   TriGenero = fem: Female (272.0/103.0)
|   |   |   |   TriGenero = mal
|   |   |   |   |   TriEdad = Bajo: Female (20.0/7.0)
|   |   |   |   |   TriEdad = Medio: Male (250.0/101.0)
|   |   |   |   |   TriEdad = Alto
```

```
|   |   |   |   |   |   RiquezaVocabulario = Pobre: Female (4.0/1.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio: Female (8.0/3.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   ErrorEdad = E-10s: Female (5.0/1.0)
|   |   |   |   |   |   |   ErrorEdad = E-20s
|   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: Male (31.0/9.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Baja: Female (2.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Alta: Male (0.0)
|   |   |   |   |   |   |   ErrorEdad = E-30s
|   |   |   |   |   |   |   |   LongitudMensaje = Corto: Male (17.0/7.0)
|   |   |   |   |   |   |   |   LongitudMensaje = Medio: Female (18.0/7.0)
|   |   |   |   |   |   |   |   LongitudMensaje = Largo: Male (2.0)
|   |   |   |   |   |   |   ErrorEdad = Sin: Male (33.0/16.0)
|   |   |   |   |   TriEdad = Null
|   |   |   |   |   |   LongitudMensaje = Corto: Male (73.0/29.0)
|   |   |   |   |   |   LongitudMensaje = Medio: Female (4.0/1.0)
|   |   |   |   |   |   LongitudMensaje = Largo: Male (1.0)
|   |   |   |   TriGenero = sin
|   |   |   |   |   UsoPalabrasCerradas = Usa
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   |   |   |   Emoticones = No-Usa
|   |   |   |   |   |   |   |   |   RiquezaVocabulario = Pobre: Male (3.0/1.0)
|   |   |   |   |   |   |   |   |   RiquezaVocabulario = Promedio: Male (29.0/12.0)
|   |   |   |   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   |   |   |   ErrorEdad = E-10s: Female (26.0/10.0)
|   |   |   |   |   |   |   |   |   |   ErrorEdad = E-20s: Male (257.0/121.0)
|   |   |   |   |   |   |   |   |   |   ErrorEdad = E-30s
|   |   |   |   |   |   |   |   |   |   |   TriEdad = Bajo: Male (1.0)
|   |   |   |   |   |   |   |   |   |   |   TriEdad = Medio: Female (21.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   TriEdad = Alto: Male (13.0/5.0)
|   |   |   |   |   |   |   |   |   |   |   TriEdad = Null: Male (34.0/14.0)
|   |   |   |   |   |   |   |   |   |   ErrorEdad = Sin: Female (382.0/172.0)
|   |   |   |   |   |   |   |   Emoticones = Baja: Female (4.0/1.0)
|   |   |   |   |   |   |   |   Emoticones = Alta: Female (0.0)
|   |   |   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   |   |   |   |   RiquezaVocabulario = Pobre: Male (1.0)
|   |   |   |   |   |   |   |   |   RiquezaVocabulario = Promedio: Female (7.0/2.0)
|   |   |   |   |   |   |   |   |   RiquezaVocabulario = Rico: Male (20.0/3.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Baja: Female (4.0/1.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Alta: Male (0.0)
|   |   |   |   |   |   |   LongitudMensaje = Largo: Female (5.0/1.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   |   |   |   RiquezaVocabulario = Pobre: Female (2.0)
|   |   |   |   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   |   |   |   TriEdad = Bajo: Female (0.0)
|   |   |   |   |   |   |   |   TriEdad = Medio: Female (20.0/8.0)
|   |   |   |   |   |   |   |   TriEdad = Alto: Female (0.0)
|   |   |   |   |   |   |   |   TriEdad = Null: Male (2.0)
|   |   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   |   TriEdad = Bajo: Female (1.0)
|   |   |   |   |   |   |   |   TriEdad = Medio
|   |   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: Male (57.0/22.0)
|   |   |   |   |   |   |   |   |   ArgotsInternet = Baja: Female (8.0/3.0)
|   |   |   |   |   |   |   |   |   ArgotsInternet = Alta: Male (0.0)
|   |   |   |   |   |   |   |   TriEdad = Alto
|   |   |   |   |   |   |   |   |   ErrorEdad = E-10s: Female (1.0)
|   |   |   |   |   |   |   |   |   ErrorEdad = E-20s: Male (13.0/4.0)
|   |   |   |   |   |   |   |   |   ErrorEdad = E-30s: Female (8.0/3.0)
|   |   |   |   |   |   |   |   |   ErrorEdad = Sin: Female (4.0/2.0)
|   |   |   |   |   |   |   |   TriEdad = Null: Female (14.0/5.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male (2.0)
|   |   |   |   |   UsoPalabrasCerradas = No-Usa: Male (17.0/6.0)
```

```
|   |   |   SignosPuntuacion = No-Usa: Male (744.0/293.0)

Number of Leaves  :   160

Size of the tree :    219


Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16542               82.71   %
Incorrectly Classified Instances      3458               17.29   %
Kappa statistic                          0.6542
Mean absolute error                      0.2315
Root mean squared error                  0.3452
Relative absolute error                 46.3013 %
Root relative squared error             69.0449 %
Total Number of Instances            20000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.834     0.179      0.823     0.834     0.828      0.902    Female
                 0.821     0.166      0.831     0.821     0.826      0.902    Male
Weighted Avg.    0.827     0.173      0.827     0.827     0.827      0.902

=== Confusion Matrix ===

    a    b   <-- classified as
 8336 1664 |    a = Female
 1794 8206 |    b = Male
```

Como se observa, no hubo un incremento muy significativo comparado con el Test anterior. Ahora se muestra la prueba realizada con el algoritmo Naive Bayes, para la clasificación de Género con el mismo Test16.arff:

```
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R15-16
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
              ErrorAciertosEdadGenero
              Genero
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===
```

Naive Bayes Classifier

Attribute	Class Female (0.5)	Male (0.5)
UsoPalabrasCerradas		
Usa	9951.0	9907.0
No-Usa	51.0	95.0
[total]	10002.0	10002.0
FrecuenciaPalabrasCerradas		
Baja	4416.0	4801.0
Media	3587.0	3205.0
Alta	2000.0	1997.0
[total]	10003.0	10003.0
RiquezaVocabulario		
Pobre	1982.0	1798.0
Promedio	3157.0	2634.0
Rico	4864.0	5571.0
[total]	10003.0	10003.0
SignosPuntuacion		
Usa	9168.0	8576.0
No-Usa	834.0	1426.0
[total]	10002.0	10002.0
LongitudMensaje		
Corto	3603.0	4035.0
Medio	2388.0	2364.0
Largo	4012.0	3604.0
[total]	10003.0	10003.0
ArgotsInternet		
No-Usa	7650.0	7701.0
Baja	2122.0	2092.0
Alta	231.0	210.0
[total]	10003.0	10003.0
Emoticones		
No-Usa	9891.0	9899.0
Baja	97.0	97.0
Alta	15.0	7.0
[total]	10003.0	10003.0
TriEdad		
Bajo	830.0	694.0
Medio	6753.0	6048.0
Alto	1526.0	2061.0
Null	895.0	1201.0
[total]	10004.0	10004.0
TriEdadGenero		
TF10	1065.0	381.0
TF20	443.0	27.0
TF30	2487.0	1335.0
TM10	611.0	956.0
TM20	823.0	2023.0
TM30	100.0	600.0
Null	4478.0	4685.0
[total]	10007.0	10007.0

```
TriGenero
  fem                          3668.0  1656.0
  mal                          4102.0  6013.0
  sin                          2233.0  2334.0
  [total]                     10003.0 10003.0

ErrorGenero
  F                            6301.0   504.0
  M                             592.0  6519.0
  S                            3110.0  2980.0
  [total]                     10003.0 10003.0

ErrorEdad
  E-10s                        1272.0  1090.0
  E-20s                        5094.0  5192.0
  E-30s                        2587.0  2709.0
  Sin                          1051.0  1013.0
  [total]                     10004.0 10004.0

ErrorAciertosEdadGenero
  H-10s                         878.0  1308.0
  H-20s                          26.0   677.0
  H-30s                          91.0  1822.0
  F-10s                        2219.0  2017.0
  F-20s                        1882.0    82.0
  F-30s                        1483.0   143.0
  Sin                          3428.0  3958.0
  [total]                     10007.0 10007.0



Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16465               82.325  %
Incorrectly Classified Instances      3535               17.675  %
Kappa statistic                          0.6465
Mean absolute error                      0.2145
Root mean squared error                  0.3564
Relative absolute error                 42.8995 %
Root relative squared error             71.2894 %
Total Number of Instances            20000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.792     0.146      0.845     0.792     0.818      0.908    Female
                 0.855     0.208      0.804     0.855     0.829      0.908    Male
Weighted Avg.    0.823     0.177      0.825     0.823     0.823      0.908

=== Confusion Matrix ===

    a    b   <-- classified as
 7920 2080 |    a = Female
 1455 8545 |    b = Male
```

Continuando con las pruebas, ahora se aplicó el algoritmo Maquinas de Soporte Vectorial el cual nos da el siguiente resultado:

```
=== Run information ===
```

```
Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K
"weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R15-16
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
              ErrorAciertosEdadGenero
              Genero
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: Female, Male

BinarySMO

Machine linear: showing attribute weights, not support vectors.

         0.4446 * (normalized) UsoPalabrasCerradas
 +       0.0731 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.0728 * (normalized) FrecuenciaPalabrasCerradas=Media
 +      -0.1459 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +       0.002  * (normalized) RiquezaVocabulario=Pobre
 +      -0.0035 * (normalized) RiquezaVocabulario=Promedio
 +       0.0015 * (normalized) RiquezaVocabulario=Rico
 +       0.6696 * (normalized) SignosPuntuacion
 +      -0.0028 * (normalized) LongitudMensaje=Corto
 +       0.0015 * (normalized) LongitudMensaje=Medio
 +       0.0013 * (normalized) LongitudMensaje=Largo
 +      -0.001  * (normalized) ArgotsInternet=No-Usa
 +       0.0022 * (normalized) ArgotsInternet=Baja
 +      -0.0012 * (normalized) ArgotsInternet=Alta
 +       0.1494 * (normalized) Emoticones=No-Usa
 +       0.1547 * (normalized) Emoticones=Baja
 +      -0.304  * (normalized) Emoticones=Alta
 +      -0.0034 * (normalized) TriEdad=Bajo
 +       0.0038 * (normalized) TriEdad=Medio
 +      -0.2206 * (normalized) TriEdad=Alto
 +       0.2202 * (normalized) TriEdad=Null
 +      -0.2537 * (normalized) TriEdadGenero=TF10
 +      -1.3537 * (normalized) TriEdadGenero=TF20
 +      -0.255  * (normalized) TriEdadGenero=TF30
 +       0.1891 * (normalized) TriEdadGenero=TM10
 +       0.4099 * (normalized) TriEdadGenero=TM20
 +       1.0737 * (normalized) TriEdadGenero=TM30
 +       0.1897 * (normalized) TriEdadGenero=Null
 +      -0.1473 * (normalized) TriGenero=fem
 +       0.0741 * (normalized) TriGenero=mal
```

```
 +       0.0732 * (normalized) TriGenero=sin
 +      -1.2615 * (normalized) ErrorGenero=F
 +       1.4073 * (normalized) ErrorGenero=M
 +      -0.1458 * (normalized) ErrorGenero=S
 +       0.0037 * (normalized) ErrorEdad=E-10s
 +       0.2174 * (normalized) ErrorEdad=E-20s
 +      -0.2225 * (normalized) ErrorEdad=E-30s
 +       0.0014 * (normalized) ErrorEdad=Sin
 +      -0.0665 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0.8225 * (normalized) ErrorAciertosEdadGenero=H-20s
 +       1.052  * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -0.0594 * (normalized) ErrorAciertosEdadGenero=F-10s
 +      -0.9555 * (normalized) ErrorAciertosEdadGenero=F-20s
 +      -0.7321 * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.061  * (normalized) ErrorAciertosEdadGenero=Sin
 -       0.3865

Number of kernel evaluations: 813108670 (23.655% cached)



Time taken to build model: 341.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16409               82.045  %
Incorrectly Classified Instances      3591               17.955  %
Kappa statistic                          0.6409
Mean absolute error                      0.1795
Root mean squared error                  0.4237
Relative absolute error                 35.91   %
Root relative squared error             84.7467 %
Total Number of Instances            20000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.814     0.173      0.825     0.814     0.819      0.82     Female
                 0.827     0.186      0.816     0.827     0.822      0.82     Male
Weighted Avg.    0.82      0.18       0.821     0.82      0.82       0.82

=== Confusion Matrix ===
    a    b   <-- classified as
 8141 1859 |    a = Female
 1732 8268 |    b = Male
```

Con estos resultados, se procede a hacer una tabla comparativa entre los tres algoritmos para identificar al clasificador que nos aporta un mayor porcentaje de efectividad y en el menor tiempo.

CLASIFICADOR	PREDICCIÓN DE GÉNERO	TIEMPO DE CONSTRUCCIÓN DEL MODELO
J48 (C4.5)	**82.71 %**	0.11 seg.
Naive Bayes	82.325%	**0.03 seg.**
Maquinas de Soporte Vectorial	82.045%	341.03 seg.

3.4.2. EDAD

Ahora, utilizamos el algoritmo J48 para identificar los tres grupos de la clase Edad, mostrando a continuación los resultados obtenidos.

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R14,16
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
              ErrorAciertosEdadGenero
              Edad
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===
J48 pruned tree
------------------
ErrorEdad = E-10s
|   TriEdad = Bajo
|   |   RiquezaVocabulario = Pobre: 10s (92.0/7.0)
|   |   RiquezaVocabulario = Promedio: 10s (122.0/23.0)
|   |   RiquezaVocabulario = Rico
|   |   |   ErrorGenero = F
|   |   |   |   ErrorAciertosEdadGenero = H-10s: 10s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = H-20s: 10s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = H-30s: 10s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-10s: 10s (37.0/7.0)
|   |   |   |   ErrorAciertosEdadGenero = F-20s: 10s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-30s: 10s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = Sin: 20s (9.0/4.0)
|   |   |   ErrorGenero = M
|   |   |   |   ErrorAciertosEdadGenero = H-10s: 10s (37.0/13.0)
|   |   |   |   ErrorAciertosEdadGenero = H-20s: 10s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = H-30s: 20s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = F-10s: 20s (11.0/3.0)
|   |   |   |   ErrorAciertosEdadGenero = F-20s: 10s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-30s: 10s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = Sin: 10s (15.0/4.0)
|   |   |   ErrorGenero = S: 20s (55.0/25.0)
|   TriEdad = Medio
|   |   TriEdadGenero = TF10
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   LongitudMensaje = Corto: 10s (4.0/1.0)
|   |   |   |   LongitudMensaje = Medio: 10s (3.0/2.0)
|   |   |   |   LongitudMensaje = Largo: 20s (12.0/3.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: 20s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: 20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   ErrorGenero = F
```

```
|   |   |   |   |   TriGenero = fem
|   |   |   |   |   |   RiquezaVocabulario = Pobre: 10s (33.0/16.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: 10s (2.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: 20s (14.0/6.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta
|   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: 20s (4.0/1.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Baja: 10s (3.0/1.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Alta: 20s (0.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: 20s (6.0/2.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: 30s (6.0/3.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (0.0)
|   |   |   |   |   TriGenero = mal: 10s (15.0/6.0)
|   |   |   |   |   TriGenero = sin: 10s (1.0)
|   |   |   |   ErrorGenero = M: 20s (16.0/4.0)
|   |   |   |   ErrorGenero = S
|   |   |   |   |   ArgotsInternet = No-Usa: 10s (11.0/4.0)
|   |   |   |   |   ArgotsInternet = Baja: 20s (5.0/2.0)
|   |   |   |   |   ArgotsInternet = Alta: 20s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: 20s (2.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: 20s (0.0)
|   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   ErrorGenero = F: 20s (3.0/1.0)
|   |   |   |   ErrorGenero = M: 10s (2.0)
|   |   |   |   ErrorGenero = S
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: 20s (2.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: 30s (4.0/1.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (0.0)
|   |   TriEdadGenero = TF20: 20s (27.0/2.0)
|   |   TriEdadGenero = TF30: 20s (414.0/145.0)
|   |   TriEdadGenero = TM10
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   ErrorGenero = F: 20s (10.0/3.0)
|   |   |   |   ErrorGenero = M: 10s (35.0/15.0)
|   |   |   |   ErrorGenero = S: 20s (6.0/3.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: 20s (2.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: 20s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: 10s (3.0/1.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: 30s (11.0/4.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (4.0)
|   |   |   |   ArgotsInternet = Baja: 20s (22.0/9.0)
|   |   |   |   ArgotsInternet = Alta: 10s (3.0/2.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: 20s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: 20s (0.0)
|   |   |   ErrorAciertosEdadGenero = Sin: 20s (23.0/9.0)
|   |   TriEdadGenero = TM20: 20s (173.0/41.0)
|   |   TriEdadGenero = TM30: 30s (17.0/4.0)
|   |   TriEdadGenero = Null
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   TriGenero = fem: 20s (19.0/5.0)
|   |   |   |   TriGenero = mal
|   |   |   |   |   ErrorGenero = F: 20s (23.0/9.0)
|   |   |   |   |   ErrorGenero = M
|   |   |   |   |   |   RiquezaVocabulario = Pobre: 20s (5.0/2.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   |   |   LongitudMensaje = Corto: 30s (0.0)
|   |   |   |   |   |   |   LongitudMensaje = Medio: 10s (3.0)
|   |   |   |   |   |   |   LongitudMensaje = Largo: 30s (12.0/4.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico: 20s (23.0/11.0)
|   |   |   |   |   ErrorGenero = S
|   |   |   |   |   |   LongitudMensaje = Corto: 30s (8.0/3.0)
```

```
|   |   |   |   |   |   LongitudMensaje = Medio: 20s (9.0/2.0)
|   |   |   |   |   |   LongitudMensaje = Largo: 10s (5.0/2.0)
|   |   |   |   TriGenero = sin
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   |   RiquezaVocabulario = Pobre: 30s (0.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio: 20s (3.0/1.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   ErrorGenero = F: 20s (3.0/1.0)
|   |   |   |   |   |   |   ErrorGenero = M
|   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: 10s (8.0/3.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Baja: 30s (2.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Alta: 10s (0.0)
|   |   |   |   |   |   |   ErrorGenero = S: 30s (14.0/5.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   |   |   LongitudMensaje = Corto: 30s (3.0)
|   |   |   |   |   |   LongitudMensaje = Medio: 10s (7.0/1.0)
|   |   |   |   |   |   LongitudMensaje = Largo: 30s (4.0/1.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 10s (2.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: 20s (3.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: 20s (2.0/1.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   ErrorGenero = F
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   |   RiquezaVocabulario = Pobre: 10s (2.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio: 10s (12.0/2.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico: 20s (29.0/15.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   |   |   LongitudMensaje = Corto: 30s (4.0)
|   |   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   |   RiquezaVocabulario = Pobre: 30s (1.0)
|   |   |   |   |   |   |   RiquezaVocabulario = Promedio: 10s (4.0/2.0)
|   |   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: 30s (22.0/11.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Baja: 20s (2.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Alta: 20s (0.0)
|   |   |   |   |   |   LongitudMensaje = Largo: 20s (55.0/29.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (54.0/25.0)
|   |   |   |   ErrorGenero = M: 20s (51.0/16.0)
|   |   |   |   ErrorGenero = S
|   |   |   |   |   SignosPuntuacion = Usa: 20s (112.0/47.0)
|   |   |   |   |   SignosPuntuacion = No-Usa: 30s (9.0/4.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: 20s (2.0/1.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: 30s (11.0/2.0)
|   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   ErrorGenero = F: 20s (9.0/3.0)
|   |   |   |   |   ErrorGenero = M: 10s (19.0/9.0)
|   |   |   |   |   ErrorGenero = S: 20s (32.0/10.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   |   RiquezaVocabulario = Pobre: 30s (2.0)
|   |   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   |   TriGenero = fem: 20s (4.0/1.0)
|   |   |   |   |   |   TriGenero = mal: 20s (12.0/3.0)
|   |   |   |   |   |   TriGenero = sin: 30s (2.0)
|   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   ArgotsInternet = No-Usa: 30s (24.0/9.0)
|   |   |   |   |   |   ArgotsInternet = Baja: 20s (4.0/1.0)
|   |   |   |   |   |   ArgotsInternet = Alta: 30s (0.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (14.0/4.0)
|   TriEdad = Alto
|   |   TriEdadGenero = TF10
|   |   |   ErrorAciertosEdadGenero = H-10s: 20s (4.0/2.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: 10s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: 10s (0.0)
```

```
|   |   |   ErrorAciertosEdadGenero = F-10s: 10s (3.0/1.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: 10s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: 10s (0.0)
|   |   |   ErrorAciertosEdadGenero = Sin: 10s (0.0)
|   |   TriEdadGenero = TF20: 20s (4.0)
|   |   TriEdadGenero = TF30: 30s (27.0/9.0)
|   |   TriEdadGenero = TM10
|   |   |   LongitudMensaje = Corto: 10s (4.0/1.0)
|   |   |   LongitudMensaje = Medio: 30s (4.0/1.0)
|   |   |   LongitudMensaje = Largo
|   |   |   |   RiquezaVocabulario = Pobre: 20s (4.0/1.0)
|   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   ErrorGenero = F: 20s (3.0/1.0)
|   |   |   |   |   ErrorGenero = M
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: 10s (0.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: 30s (3.0/1.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 10s (8.0/4.0)
|   |   |   |   |   ErrorGenero = S: 30s (1.0)
|   |   |   |   RiquezaVocabulario = Rico: 20s (4.0/1.0)
|   |   TriEdadGenero = TM20: 20s (27.0/7.0)
|   |   TriEdadGenero = TM30: 30s (14.0/3.0)
|   |   TriEdadGenero = Null
|   |   |   SignosPuntuacion = Usa: 30s (80.0/35.0)
|   |   |   SignosPuntuacion = No-Usa
|   |   |   |   ErrorGenero = F
|   |   |   |   |   TriGenero = fem: 20s (0.0)
|   |   |   |   |   TriGenero = mal: 20s (2.0)
|   |   |   |   |   TriGenero = sin: 30s (2.0)
|   |   |   |   ErrorGenero = M: 10s (12.0/6.0)
|   |   |   |   ErrorGenero = S: 20s (11.0/3.0)
|   TriEdad = Null
|   |   RiquezaVocabulario = Pobre: 10s (1.0)
|   |   RiquezaVocabulario = Promedio: 10s (2.0)
|   |   RiquezaVocabulario = Rico
|   |   |   ErrorAciertosEdadGenero = H-10s: 30s (56.0/31.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: 20s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: 20s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   SignosPuntuacion = Usa
|   |   |   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   |   |   TriGenero = fem: 20s (7.0/3.0)
|   |   |   |   |   |   |   TriGenero = mal: 30s (8.0/2.0)
|   |   |   |   |   |   |   TriGenero = sin: 20s (18.0/4.0)
|   |   |   |   |   |   LongitudMensaje = Medio: 30s (2.0)
|   |   |   |   |   |   LongitudMensaje = Largo: 20s (2.0/1.0)
|   |   |   |   |   SignosPuntuacion = No-Usa
|   |   |   |   |   |   ErrorGenero = F: 10s (3.0/1.0)
|   |   |   |   |   |   ErrorGenero = M: 20s (9.0/4.0)
|   |   |   |   |   |   ErrorGenero = S: 30s (26.0/15.0)
|   |   |   |   ArgotsInternet = Baja
|   |   |   |   |   SignosPuntuacion = Usa: 30s (6.0/2.0)
|   |   |   |   |   SignosPuntuacion = No-Usa: 20s (7.0)
|   |   |   |   ArgotsInternet = Alta: 20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: 30s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: 30s (1.0)
|   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   ErrorGenero = F
|   |   |   |   |   TriGenero = fem: 30s (0.0)
|   |   |   |   |   TriGenero = mal: 20s (2.0)
|   |   |   |   |   TriGenero = sin: 30s (6.0/1.0)
|   |   |   |   ErrorGenero = M: 20s (14.0/5.0)
|   |   |   |   ErrorGenero = S
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: 20s (38.0/16.0)
```

```
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: 30s (2.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (0.0)
ErrorEdad = E-20s
|   ErrorGenero = F
|   |   ErrorAciertosEdadGenero = H-10s: 20s (221.0/23.0)
|   |   ErrorAciertosEdadGenero = H-20s: 20s (9.0/2.0)
|   |   ErrorAciertosEdadGenero = H-30s
|   |   |   TriEdad = Bajo: 20s (0.0)
|   |   |   TriEdad = Medio: 20s (12.0/2.0)
|   |   |   TriEdad = Alto: 30s (4.0)
|   |   |   TriEdad = Null: 20s (0.0)
|   |   ErrorAciertosEdadGenero = F-10s: 20s (763.0/87.0)
|   |   ErrorAciertosEdadGenero = F-20s: 20s (1728.0/16.0)
|   |   ErrorAciertosEdadGenero = F-30s
|   |   |   ArgotsInternet = No-Usa: 30s (51.0/15.0)
|   |   |   ArgotsInternet = Baja
|   |   |   |   LongitudMensaje = Corto: 20s (1.0)
|   |   |   |   LongitudMensaje = Medio: 30s (2.0)
|   |   |   |   LongitudMensaje = Largo: 20s (14.0/6.0)
|   |   |   ArgotsInternet = Alta: 20s (1.0)
|   |   ErrorAciertosEdadGenero = Sin: 20s (919.0/144.0)
|   ErrorGenero = M: 20s (3774.0/255.0)
|   ErrorGenero = S
|   |   TriEdad = Bajo: 20s (225.0/85.0)
|   |   TriEdad = Medio: 20s (1759.0/444.0)
|   |   TriEdad = Alto
|   |   |   ArgotsInternet = No-Usa
|   |   |   |   TriEdadGenero = TF10: 20s (2.0/1.0)
|   |   |   |   TriEdadGenero = TF20: 30s (0.0)
|   |   |   |   TriEdadGenero = TF30: 30s (30.0/8.0)
|   |   |   |   TriEdadGenero = TM10
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: 20s (10.0/3.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: 30s (3.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (0.0)
|   |   |   |   TriEdadGenero = TM20
|   |   |   |   |   LongitudMensaje = Corto: 20s (9.0/4.0)
|   |   |   |   |   LongitudMensaje = Medio: 30s (2.0)
|   |   |   |   |   LongitudMensaje = Largo: 30s (0.0)
|   |   |   |   TriEdadGenero = TM30: 30s (14.0/2.0)
|   |   |   |   TriEdadGenero = Null
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   |   ErrorAciertosEdadGenero = H-10s: 20s (10.0/3.0)
|   |   |   |   |   |   ErrorAciertosEdadGenero = H-20s: 20s (1.0)
|   |   |   |   |   |   ErrorAciertosEdadGenero = H-30s: 30s (8.0/1.0)
|   |   |   |   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   |   |   |   SignosPuntuacion = Usa: 30s (11.0/4.0)
|   |   |   |   |   |   |   SignosPuntuacion = No-Usa: 20s (6.0/1.0)
|   |   |   |   |   |   ErrorAciertosEdadGenero = F-20s: 20s (3.0)
|   |   |   |   |   |   ErrorAciertosEdadGenero = F-30s: 20s (2.0/1.0)
|   |   |   |   |   |   ErrorAciertosEdadGenero = Sin: 30s (85.0/31.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: 20s (41.0/12.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: 30s (0.0)
|   |   |   ArgotsInternet = Baja: 20s (18.0/4.0)
|   |   |   ArgotsInternet = Alta: 30s (0.0)
|   |   TriEdad = Null
|   |   |   UsoPalabrasCerradas = Usa: 20s (534.0/231.0)
|   |   |   UsoPalabrasCerradas = No-Usa
|   |   |   |   RiquezaVocabulario = Pobre: 20s (4.0/2.0)
|   |   |   |   RiquezaVocabulario = Promedio: 10s (2.0/1.0)
|   |   |   |   RiquezaVocabulario = Rico: 30s (6.0/1.0)
ErrorEdad = E-30s: 30s (5294.0/454.0)
ErrorEdad = Sin
|   TriEdadGenero = TF10
|   |   ErrorAciertosEdadGenero = H-10s: 30s (5.0)
```

```
|   |   ErrorAciertosEdadGenero = H-20s: 20s (0.0)
|   |   ErrorAciertosEdadGenero = H-30s: 20s (0.0)
|   |   ErrorAciertosEdadGenero = F-10s
|   |   |   FrecuenciaPalabrasCerradas = Baja: 30s (5.0/2.0)
|   |   |   FrecuenciaPalabrasCerradas = Media: 20s (3.0/1.0)
|   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (0.0)
|   |   ErrorAciertosEdadGenero = F-20s: 20s (0.0)
|   |   ErrorAciertosEdadGenero = F-30s: 20s (3.0/1.0)
|   |   ErrorAciertosEdadGenero = Sin
|   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   ErrorGenero = F: 20s (7.0/1.0)
|   |   |   |   ErrorGenero = M: 30s (2.0)
|   |   |   |   ErrorGenero = S
|   |   |   |   |   RiquezaVocabulario = Pobre: 20s (12.0/4.0)
|   |   |   |   |   RiquezaVocabulario = Promedio: 30s (17.0/8.0)
|   |   |   |   |   RiquezaVocabulario = Rico: 30s (25.0/10.0)
|   |   |   FrecuenciaPalabrasCerradas = Media: 20s (4.0)
|   |   |   FrecuenciaPalabrasCerradas = Alta: 20s (0.0)
|   TriEdadGenero = TF20: 20s (14.0)
|   TriEdadGenero = TF30
|   |   TriEdad = Bajo: 20s (18.0/7.0)
|   |   TriEdad = Medio: 20s (327.0/136.0)
|   |   TriEdad = Alto
|   |   |   ArgotsInternet = No-Usa: 30s (35.0/12.0)
|   |   |   ArgotsInternet = Baja: 20s (2.0)
|   |   |   ArgotsInternet = Alta: 30s (0.0)
|   |   TriEdad = Null: 20s (0.0)
|   TriEdadGenero = TM10
|   |   ErrorGenero = F: 20s (4.0/2.0)
|   |   ErrorGenero = M: 30s (7.0/2.0)
|   |   ErrorGenero = S: 20s (29.0/12.0)
|   TriEdadGenero = TM20: 20s (54.0/17.0)
|   TriEdadGenero = TM30: 30s (17.0)
|   TriEdadGenero = Null
|   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   Emoticones = No-Usa: 30s (1331.0/565.0)
|   |   |   Emoticones = Baja: 20s (2.0)
|   |   |   Emoticones = Alta: 30s (0.0)
|   |   FrecuenciaPalabrasCerradas = Media
|   |   |   TriGenero = fem: 20s (30.0/9.0)
|   |   |   TriGenero = mal: 20s (47.0/21.0)
|   |   |   TriGenero = sin
|   |   |   |   RiquezaVocabulario = Pobre: 20s (2.0)
|   |   |   |   RiquezaVocabulario = Promedio: 20s (13.0/4.0)
|   |   |   |   RiquezaVocabulario = Rico: 30s (43.0/16.0)
|   |   FrecuenciaPalabrasCerradas = Alta: 20s (4.0/1.0)

Number of Leaves  :   242
Size of the tree :   335

Time taken to build model: 0.34 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16218               81.09   %
Incorrectly Classified Instances      3782               18.91   %
Kappa statistic                          0.6273
Mean absolute error                      0.1749
Root mean squared error                  0.3013
Relative absolute error                 50.6242 %
Root relative squared error             72.4927 %
Total Number of Instances            20000
```

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.442     0.008      0.662     0.442     0.53       0.913    10s
                 0.874     0.257      0.817     0.874     0.844      0.887    20s
                 0.753     0.116      0.811     0.753     0.781      0.889    30s
Weighted Avg.    0.811     0.193      0.809     0.811     0.808      0.889

=== Confusion Matrix ===

    a    b    c   <-- classified as
  303  310   72 |    a = 10s
  104 9916 1328 |    b = 20s
   51 1917 5999 |    c = 30s
```

Continuando ahora con el resultado obtenido con el algoritmo Naive Bayes. Los resultados fueron los siguientes:

```
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R14,16
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
              ErrorAciertosEdadGenero
              Edad
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                               Class
Attribute                        10s     20s     30s
                               (0.03)  (0.57)   (0.4)
=====================================================
UsoPalabrasCerradas
  Usa                           674.0 11278.0  7907.0
  No-Usa                         13.0    72.0    62.0
  [total]                       687.0 11350.0  7969.0

FrecuenciaPalabrasCerradas
  Baja                          400.0  5100.0  3718.0
  Media                         193.0  4034.0  2566.0
  Alta                           95.0  2217.0  1686.0
  [total]                       688.0 11351.0  7970.0

RiquezaVocabulario
  Pobre                         177.0  2442.0  1162.0
  Promedio                      229.0  3566.0  1997.0
```

```
  Rico                          282.0  5343.0  4811.0
  [total]                       688.0 11351.0  7970.0

SignosPuntuacion
  Usa                           582.0 10168.0  6995.0
  No-Usa                        105.0  1182.0   974.0
  [total]                       687.0 11350.0  7969.0

LongitudMensaje
  Corto                         309.0  4028.0  3302.0
  Medio                         187.0  2834.0  1732.0
  Largo                         192.0  4489.0  2936.0
  [total]                       688.0 11351.0  7970.0

ArgotsInternet
  No-Usa                        528.0  8468.0  6356.0
  Baja                          147.0  2601.0  1467.0
  Alta                           13.0   282.0   147.0
  [total]                       688.0 11351.0  7970.0

Emoticones
  No-Usa                        669.0 11167.0  7955.0
  Baja                           14.0   167.0    14.0
  Alta                            5.0    17.0     1.0
  [total]                       688.0 11351.0  7970.0

TriEdad
  Bajo                          295.0   736.0   494.0
  Medio                         298.0  8496.0  4008.0
  Alto                           34.0  1127.0  2427.0
  Null                           62.0   993.0  1042.0
  [total]                       689.0 11352.0  7971.0

TriEdadGenero
  TF10                          194.0   875.0   378.0
  TF20                            1.0   460.0    10.0
  TF30                           70.0  2267.0  1486.0
  TM10                          156.0   749.0   663.0
  TM20                           16.0  2132.0   699.0
  TM30                            1.0    78.0   622.0
  Null                          254.0  4794.0  4116.0
  [total]                       692.0 11355.0  7974.0

TriGenero
  fem                           255.0  3548.0  1522.0
  mal                           277.0  5495.0  4344.0
  sin                           156.0  2308.0  2104.0
  [total]                       688.0 11351.0  7970.0

ErrorGenero
  F                             249.0  4047.0  2510.0
  M                             245.0  4084.0  2783.0
  S                             194.0  3220.0  2677.0
  [total]                       688.0 11351.0  7970.0

ErrorEdad
  E-10s                         573.0  1153.0   637.0
  E-20s                          71.0  8804.0  1412.0
  E-30s                           4.0   452.0  4841.0
  Sin                            41.0   943.0  1081.0
  [total]                       689.0 11352.0  7971.0

ErrorAciertosEdadGenero
  H-10s                         217.0  1254.0   716.0
```

```
  H-20s                          1.0   681.0    22.0
  H-30s                          1.0   172.0  1741.0
  F-10s                        332.0  2825.0  1080.0
  F-20s                          1.0  1930.0    34.0
  F-30s                          1.0   160.0  1466.0
  Sin                          139.0  4333.0  2915.0
  [total]                      692.0 11355.0  7974.0


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16134               80.67   %
Incorrectly Classified Instances      3866               19.33   %
Kappa statistic                          0.6248
Mean absolute error                      0.1597
Root mean squared error                  0.3053
Relative absolute error                 46.2359 %
Root relative squared error             73.4599 %
Total Number of Instances            20000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.533     0.018      0.506     0.533     0.519      0.949    10s
                 0.857     0.238      0.825     0.857     0.841      0.9      20s
                 0.759     0.121      0.806     0.759     0.782      0.903    30s
Weighted Avg.    0.807     0.184      0.807     0.807     0.806      0.903

=== Confusion Matrix ===

    a    b    c   <-- classified as
  365  258   62 |    a = 10s
  230 9726 1392 |    b = 20s
  126 1798 6043 |    c = 30s
```

Por ultimo para la clasificación de Edad se utilizó el algoritmo Maquinas de Soporte Vectorial. Los resultados fueron los siguientes:

```
=== Run information ===

Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K
"weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R14,16
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
```

```
              ErrorAciertosEdadGenero
              Edad
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: 10s, 20s

BinarySMO

Machine linear: showing attribute weights, not support vectors.

        -1.9955 * (normalized) UsoPalabrasCerradas
 +      -0.0002 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.0001 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.0002 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.0005 * (normalized) RiquezaVocabulario=Pobre
 +      -0.0002 * (normalized) RiquezaVocabulario=Promedio
 +       0.0006 * (normalized) RiquezaVocabulario=Rico
 +      -0.0006 * (normalized) SignosPuntuacion
 +      -0.0002 * (normalized) LongitudMensaje=Corto
 +      -0.0001 * (normalized) LongitudMensaje=Medio
 +       0.0003 * (normalized) LongitudMensaje=Largo
 +      -0.0002 * (normalized) ArgotsInternet=No-Usa
 +      -0.0002 * (normalized) ArgotsInternet=Baja
 +       0.0003 * (normalized) ArgotsInternet=Alta
 +       0.6658 * (normalized) Emoticones=No-Usa
 +       0.665  * (normalized) Emoticones=Baja
 +      -1.3308 * (normalized) Emoticones=Alta
 +      -1.4986 * (normalized) TriEdad=Bajo
 +       0.4995 * (normalized) TriEdad=Medio
 +       0.4996 * (normalized) TriEdad=Alto
 +       0.4995 * (normalized) TriEdad=Null
 +      -0.0014 * (normalized) TriEdadGenero=TF10
 +       0.001  * (normalized) TriEdadGenero=TF20
 +       0.0001 * (normalized) TriEdadGenero=TF30
 +      -0.0014 * (normalized) TriEdadGenero=TM10
 +       0.0004 * (normalized) TriEdadGenero=TM20
 +       0.0017 * (normalized) TriEdadGenero=TM30
 +      -0.0005 * (normalized) TriEdadGenero=Null
 +       0.0003 * (normalized) TriGenero=fem
 +       0      * (normalized) TriGenero=mal
 +      -0.0003 * (normalized) TriGenero=sin
 +      -0.0001 * (normalized) ErrorGenero=F
 +      -0.0003 * (normalized) ErrorGenero=M
 +       0.0004 * (normalized) ErrorGenero=S
 +      -1.4988 * (normalized) ErrorEdad=E-10s
 +       0.5004 * (normalized) ErrorEdad=E-20s
 +       0.4997 * (normalized) ErrorEdad=E-30s
 +       0.4987 * (normalized) ErrorEdad=Sin
 +      -0.7496 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-20s
 +       1      * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -0.7493 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       1.2479 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.749  * (normalized) ErrorAciertosEdadGenero=Sin
 +       2.0831
```

```
Number of kernel evaluations: 71227693 (39.51% cached)

Classifier for classes: 10s, 30s

BinarySMO

Machine linear: showing attribute weights, not support vectors.

        -0.4999 * (normalized) UsoPalabrasCerradas
 +      -0.3338 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.0422 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.2916 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.5002 * (normalized) RiquezaVocabulario=Pobre
 +      -0.1251 * (normalized) RiquezaVocabulario=Promedio
 +       0.6253 * (normalized) RiquezaVocabulario=Rico
 +      -0.4993 * (normalized) SignosPuntuacion
 +      -0.0836 * (normalized) LongitudMensaje=Corto
 +      -0.2081 * (normalized) LongitudMensaje=Medio
 +       0.2917 * (normalized) LongitudMensaje=Largo
 +       0.2502 * (normalized) ArgotsInternet=No-Usa
 +       0.1247 * (normalized) ArgotsInternet=Baja
 +      -0.3749 * (normalized) ArgotsInternet=Alta
 +       1.1255 * (normalized) Emoticones=No-Usa
 +      -0.249  * (normalized) Emoticones=Baja
 +      -0.8765 * (normalized) Emoticones=Alta
 +      -0.5936 * (normalized) TriEdad=Bajo
 +       0.0324 * (normalized) TriEdad=Medio
 +       0.4046 * (normalized) TriEdad=Alto
 +       0.1566 * (normalized) TriEdad=Null
 +      -0.7853 * (normalized) TriEdadGenero=TF10
 +       0.2135 * (normalized) TriEdadGenero=TF20
 +       0.2137 * (normalized) TriEdadGenero=TF30
 +      -1.0356 * (normalized) TriEdadGenero=TM10
 +       0.3364 * (normalized) TriEdadGenero=TM20
 +       1.0928 * (normalized) TriEdadGenero=TM30
 +      -0.0357 * (normalized) TriEdadGenero=Null
 +      -0.1248 * (normalized) TriGenero=fem
 +       0.1252 * (normalized) TriGenero=mal
 +      -0.0004 * (normalized) TriGenero=sin
 +      -0.0422 * (normalized) ErrorGenero=F
 +      -0.2911 * (normalized) ErrorGenero=M
 +       0.3333 * (normalized) ErrorGenero=S
 +      -1.1875 * (normalized) ErrorEdad=E-10s
 +       0.063  * (normalized) ErrorEdad=E-20s
 +       1.1869 * (normalized) ErrorEdad=E-30s
 +      -0.0624 * (normalized) ErrorEdad=Sin
 +      -0.9574 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-20s
 +       0.7902 * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -0.8328 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0.5422 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0.9155 * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.4577 * (normalized) ErrorAciertosEdadGenero=Sin
 +       1.1069

Number of kernel evaluations: 21105718 (51.306% cached)

Classifier for classes: 20s, 30s

BinarySMO

Machine linear: showing attribute weights, not support vectors.

        -0.0005 * (normalized) UsoPalabrasCerradas
```

```
 +       0      * (normalized) FrecuenciaPalabrasCerradas=Baja
 +      -0.0001 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.0001 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.0001 * (normalized) RiquezaVocabulario=Pobre
 +       0      * (normalized) RiquezaVocabulario=Promedio
 +       0.0001 * (normalized) RiquezaVocabulario=Rico
 +      -0.0001 * (normalized) SignosPuntuacion
 +       0.0001 * (normalized) LongitudMensaje=Corto
 +      -0.0001 * (normalized) LongitudMensaje=Medio
 +       0      * (normalized) LongitudMensaje=Largo
 +       0.0001 * (normalized) ArgotsInternet=No-Usa
 +       0      * (normalized) ArgotsInternet=Baja
 +       0      * (normalized) ArgotsInternet=Alta
 +       0      * (normalized) Emoticones=No-Usa
 +       0      * (normalized) Emoticones=Baja
 +      -0.0001 * (normalized) TriEdad=Bajo
 +      -0.0001 * (normalized) TriEdad=Medio
 +       0.0002 * (normalized) TriEdad=Alto
 +       0.0001 * (normalized) TriEdad=Null
 +       0.2855 * (normalized) TriEdadGenero=TF10
 +      -1.7148 * (normalized) TriEdadGenero=TF20
 +       0.2857 * (normalized) TriEdadGenero=TF30
 +       0.2856 * (normalized) TriEdadGenero=TM10
 +       0.2854 * (normalized) TriEdadGenero=TM20
 +       0.287  * (normalized) TriEdadGenero=TM30
 +       0.2857 * (normalized) TriEdadGenero=Null
 +       0      * (normalized) TriGenero=fem
 +      -0.0001 * (normalized) TriGenero=mal
 +       0.0001 * (normalized) TriGenero=sin
 +      -0.0001 * (normalized) ErrorGenero=F
 +      -0.0001 * (normalized) ErrorGenero=M
 +       0.0002 * (normalized) ErrorGenero=S
 +      -0.9996 * (normalized) ErrorEdad=E-10s
 +      -0.9998 * (normalized) ErrorEdad=E-20s
 +       1.0002 * (normalized) ErrorEdad=E-30s
 +       0.9992 * (normalized) ErrorEdad=Sin
 +       0.0002 * (normalized) ErrorAciertosEdadGenero=H-10s
 +      -0.0006 * (normalized) ErrorAciertosEdadGenero=H-20s
 +       0.0005 * (normalized) ErrorAciertosEdadGenero=H-30s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-10s
 +      -0.0008 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0.0006 * (normalized) ErrorAciertosEdadGenero=F-30s
 +       0.0001 * (normalized) ErrorAciertosEdadGenero=Sin
 -       0.2859

Number of kernel evaluations: 552184101 (20.2% cached)



Time taken to build model: 257.28 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16076               80.38   %
Incorrectly Classified Instances      3924               19.62   %
Kappa statistic                          0.6117
Mean absolute error                      0.2703
Root mean squared error                  0.3491
Relative absolute error                 78.238  %
Root relative squared error             83.9989 %
Total Number of Instances            20000

=== Detailed Accuracy By Class ===
```

```
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.391     0.007      0.654     0.391     0.489      0.783    10s
                   0.872     0.276      0.806     0.872     0.838      0.798    20s
                   0.742     0.116      0.809     0.742     0.774      0.815    30s
Weighted Avg.      0.804     0.203      0.802     0.804     0.8        0.804

=== Confusion Matrix ===

    a    b    c   <-- classified as
  268  375   42 |    a = 10s
   97 9898 1353 |    b = 20s
   45 2012 5910 |    c = 30s
```

Con estos resultados, se procede a hacer una tabla comparativa entre los tres algoritmos para identificar al clasificador que nos aporta un mayor porcentaje de efectividad en el menor tiempo.

Tabla 1. Comparativa de efectividad por Clasificador (para Género).

CLASIFICADOR	PREDICCIÓN DE EDAD	TIEMPO DE CONSTRUCCIÓN DEL MODELO
J48 (C4.5)	**81.09 %**	0.34 seg.
Naive Bayes	80.67 %	**0.03 seg.**
Maquinas de Soporte Vectorial	80.38 %	257.28 seg.

3.4.3. GÉNERO / EDAD

Ahora se realizaron las mismas pruebas con los tres algoritmos utilizados anterior mente, para clasificar la combinación de las clases edad y género de manera conjunta.

Primero comenzaremos implementando el algoritmo J48 y a continuación mostramos los resultados obtenidos.

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R14-15
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
              ErrorAciertosEdadGenero
              General
Test mode:10-fold cross-validation
```

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

ErrorGenero = F
|   ErrorEdad = E-10s
|   |   TriEdad = Bajo
|   |   |   ErrorAciertosEdadGenero = H-10s: Female-20s (3.0/1.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Female-10s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Female-10s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-10s: Female-10s (125.0/12.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: Female-10s (0.0)
|   |   |   ErrorAciertosEdadGenero = Sin: Female-20s (10.0/5.0)
|   |   TriEdad = Medio
|   |   |   ErrorAciertosEdadGenero = H-10s: Female-20s (91.0/41.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Female-20s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   TriEdadGenero = TF10
|   |   |   |   |   TriGenero = fem
|   |   |   |   |   |   RiquezaVocabulario = Pobre: Female-10s (33.0/16.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female-10s (2.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-20s (14.0/7.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta
|   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: Female-20s (4.0/2.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Baja: Female-10s (3.0/1.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Alta: Female-20s (0.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female-20s (6.0/2.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-30s (6.0/3.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (0.0)
|   |   |   |   |   TriGenero = mal: Female-10s (15.0/6.0)
|   |   |   |   |   TriGenero = sin: Female-10s (1.0)
|   |   |   |   TriEdadGenero = TF20: Female-20s (8.0/1.0)
|   |   |   |   TriEdadGenero = TF30
|   |   |   |   |   LongitudMensaje = Corto: Female-10s (16.0/9.0)
|   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   RiquezaVocabulario = Pobre: Female-20s (6.0/3.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio: Female-20s (14.0/5.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   TriGenero = fem: Female-30s (10.0/4.0)
|   |   |   |   |   |   |   TriGenero = mal: Female-10s (2.0/1.0)
|   |   |   |   |   |   |   TriGenero = sin: Female-20s (2.0)
|   |   |   |   |   LongitudMensaje = Largo: Female-20s (57.0/23.0)
|   |   |   |   TriEdadGenero = TM10
|   |   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female-10s (3.0/1.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-30s (8.0/3.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (4.0/1.0)
|   |   |   |   |   ArgotsInternet = Baja: Female-20s (17.0/10.0)
|   |   |   |   |   ArgotsInternet = Alta: Female-10s (1.0)
|   |   |   |   TriEdadGenero = TM20
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female-20s (5.0/2.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-20s (9.0/3.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (20.0/9.0)
|   |   |   |   TriEdadGenero = TM30: Male-20s (2.0/1.0)
|   |   |   |   TriEdadGenero = Null
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   |   RiquezaVocabulario = Pobre: Female-10s (2.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio: Female-10s (12.0/2.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico: Female-20s (29.0/17.0)
```

```
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   |   |   RiquezaVocabulario = Pobre
|   |   |   |   |   |   |   ArgotsInternet = No-Usa: Female-10s (9.0/4.0)
|   |   |   |   |   |   |   ArgotsInternet = Baja: Female-20s (9.0/4.0)
|   |   |   |   |   |   |   ArgotsInternet = Alta: Female-10s (1.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio: Female-20s (33.0/20.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   ArgotsInternet = No-Usa: Female-30s (30.0/17.0)
|   |   |   |   |   |   |   ArgotsInternet = Baja: Female-20s (6.0/3.0)
|   |   |   |   |   |   |   ArgotsInternet = Alta: Female-30s (0.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta
|   |   |   |   |   |   ArgotsInternet = No-Usa: Female-20s (25.0/14.0)
|   |   |   |   |   |   ArgotsInternet = Baja: Female-30s (22.0/14.0)
|   |   |   |   |   |   ArgotsInternet = Alta: Female-20s (7.0/2.0)
|   |   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (8.0/1.0)
|   |   |   |   ErrorAciertosEdadGenero = F-30s: Female-30s (13.0/3.0)
|   |   |   |   ErrorAciertosEdadGenero = Sin: Female-20s (76.0/28.0)
|   |   |   TriEdad = Alto
|   |   |   |   TriEdadGenero = TF10: Female-10s (1.0)
|   |   |   |   TriEdadGenero = TF20: Female-20s (1.0)
|   |   |   |   TriEdadGenero = TF30: Female-30s (9.0/3.0)
|   |   |   |   TriEdadGenero = TM10
|   |   |   |   |   RiquezaVocabulario = Pobre: Female-10s (0.0)
|   |   |   |   |   RiquezaVocabulario = Promedio: Female-20s (3.0/1.0)
|   |   |   |   |   RiquezaVocabulario = Rico: Female-10s (2.0)
|   |   |   |   TriEdadGenero = TM20: Female-20s (4.0)
|   |   |   |   TriEdadGenero = TM30: Male-20s (2.0/1.0)
|   |   |   |   TriEdadGenero = Null
|   |   |   |   |   LongitudMensaje = Corto: Female-20s (10.0/5.0)
|   |   |   |   |   LongitudMensaje = Medio: Female-30s (9.0/3.0)
|   |   |   |   |   LongitudMensaje = Largo: Female-30s (6.0/3.0)
|   |   |   TriEdad = Null
|   |   |   |   TriGenero = fem: Female-20s (1.0)
|   |   |   |   TriGenero = mal: Female-20s (6.0/2.0)
|   |   |   |   TriGenero = sin: Female-30s (18.0/12.0)
|   |   ErrorEdad = E-20s
|   |   |   ErrorAciertosEdadGenero = H-10s: Female-20s (221.0/36.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Male-20s (9.0/5.0)
|   |   |   ErrorAciertosEdadGenero = H-30s
|   |   |   |   TriEdad = Bajo: Female-20s (0.0)
|   |   |   |   TriEdad = Medio: Female-20s (12.0/4.0)
|   |   |   |   TriEdad = Alto: Male-30s (4.0/2.0)
|   |   |   |   TriEdad = Null: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-10s: Female-20s (763.0/156.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (1728.0/24.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: Female-30s (69.0/25.0)
|   |   |   ErrorAciertosEdadGenero = Sin: Female-20s (919.0/212.0)
|   |   ErrorEdad = E-30s: Female-30s (2021.0/231.0)
|   |   ErrorEdad = Sin
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   TriEdad = Bajo: Female-20s (2.0)
|   |   |   |   TriEdad = Medio: Female-30s (4.0/1.0)
|   |   |   |   TriEdad = Alto: Female-30s (1.0)
|   |   |   |   TriEdad = Null: Female-30s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Female-30s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Male-30s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   TriEdad = Bajo: Female-20s (4.0/2.0)
|   |   |   |   TriEdad = Medio
|   |   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   |   LongitudMensaje = Corto: Female-20s (9.0/3.0)
|   |   |   |   |   |   LongitudMensaje = Medio: Female-30s (8.0/4.0)
|   |   |   |   |   |   LongitudMensaje = Largo: Female-30s (5.0/2.0)
|   |   |   |   |   ArgotsInternet = Baja: Male-20s (3.0/1.0)
```

```
|   |   |   |   ArgotsInternet = Alta: Male-20s (1.0)
|   |   |   TriEdad = Alto: Male-20s (2.0/1.0)
|   |   |   TriEdad = Null: Female-30s (7.0/3.0)
|   |   ErrorAciertosEdadGenero = F-20s: Female-20s (3.0/1.0)
|   |   ErrorAciertosEdadGenero = F-30s
|   |   |   TriEdadGenero = TF10: Male-20s (2.0/1.0)
|   |   |   TriEdadGenero = TF20: Female-30s (0.0)
|   |   |   TriEdadGenero = TF30: Male-20s (2.0)
|   |   |   TriEdadGenero = TM10: Female-30s (0.0)
|   |   |   TriEdadGenero = TM20: Female-30s (0.0)
|   |   |   TriEdadGenero = TM30: Female-30s (0.0)
|   |   |   TriEdadGenero = Null: Female-30s (10.0/4.0)
|   |   ErrorAciertosEdadGenero = Sin
|   |   |   TriEdadGenero = TF10
|   |   |   |   LongitudMensaje = Corto: Female-20s (5.0/1.0)
|   |   |   |   LongitudMensaje = Medio: Male-20s (3.0)
|   |   |   |   LongitudMensaje = Largo: Female-20s (0.0)
|   |   |   TriEdadGenero = TF20: Female-20s (2.0)
|   |   |   TriEdadGenero = TF30: Female-30s (21.0/9.0)
|   |   |   TriEdadGenero = TM10: Male-20s (3.0/2.0)
|   |   |   TriEdadGenero = TM20: Male-30s (3.0/2.0)
|   |   |   TriEdadGenero = TM30: Female-30s (0.0)
|   |   |   TriEdadGenero = Null
|   |   |   |   LongitudMensaje = Corto: Female-30s (76.0/46.0)
|   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   ArgotsInternet = No-Usa: Female-30s (13.0/7.0)
|   |   |   |   |   ArgotsInternet = Baja: Male-30s (6.0/3.0)
|   |   |   |   |   ArgotsInternet = Alta: Female-20s (0.0)
|   |   |   |   LongitudMensaje = Largo: Female-20s (8.0/4.0)
ErrorGenero = M
|   ErrorEdad = E-10s
|   |   ErrorAciertosEdadGenero = H-10s
|   |   |   TriEdadGenero = TF10: Male-10s (26.0/11.0)
|   |   |   TriEdadGenero = TF20: Female-20s (4.0/1.0)
|   |   |   TriEdadGenero = TF30: Male-10s (37.0/21.0)
|   |   |   TriEdadGenero = TM10: Male-10s (122.0/29.0)
|   |   |   TriEdadGenero = TM20: Male-20s (53.0/22.0)
|   |   |   TriEdadGenero = TM30: Male-30s (7.0/3.0)
|   |   |   TriEdadGenero = Null
|   |   |   |   TriGenero = fem: Male-10s (8.0/3.0)
|   |   |   |   TriGenero = mal
|   |   |   |   |   RiquezaVocabulario = Pobre: Male-20s (5.0/2.0)
|   |   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   |   LongitudMensaje = Corto: Male-10s (0.0)
|   |   |   |   |   |   LongitudMensaje = Medio: Male-10s (3.0)
|   |   |   |   |   |   LongitudMensaje = Largo: Male-30s (14.0/8.0)
|   |   |   |   |   RiquezaVocabulario = Rico: Male-20s (46.0/30.0)
|   |   |   |   TriGenero = sin: Male-10s (44.0/20.0)
|   |   ErrorAciertosEdadGenero = H-20s: Male-20s (6.0)
|   |   ErrorAciertosEdadGenero = H-30s: Male-20s (5.0/2.0)
|   |   ErrorAciertosEdadGenero = F-10s: Male-20s (195.0/81.0)
|   |   ErrorAciertosEdadGenero = F-20s: Female-20s (1.0)
|   |   ErrorAciertosEdadGenero = F-30s: Female-30s (1.0)
|   |   ErrorAciertosEdadGenero = Sin
|   |   |   TriEdad = Bajo: Male-10s (28.0/5.0)
|   |   |   TriEdad = Medio
|   |   |   |   TriEdadGenero = TF10: Male-10s (2.0)
|   |   |   |   TriEdadGenero = TF20: Female-20s (1.0)
|   |   |   |   TriEdadGenero = TF30
|   |   |   |   |   TriGenero = fem
|   |   |   |   |   |   LongitudMensaje = Corto: Male-30s (7.0/4.0)
|   |   |   |   |   |   LongitudMensaje = Medio: Male-20s (6.0/2.0)
|   |   |   |   |   |   LongitudMensaje = Largo: Female-20s (5.0/3.0)
|   |   |   |   |   TriGenero = mal: Male-20s (2.0/1.0)
```

```
|   |   |   |   |   TriGenero = sin: Male-10s (1.0)
|   |   |   |   TriEdadGenero = TM10
|   |   |   |   |   LongitudMensaje = Corto: Male-20s (4.0/2.0)
|   |   |   |   |   LongitudMensaje = Medio: Male-10s (2.0)
|   |   |   |   |   LongitudMensaje = Largo: Male-20s (4.0/2.0)
|   |   |   |   TriEdadGenero = TM20: Male-20s (12.0/3.0)
|   |   |   |   TriEdadGenero = TM30: Male-20s (2.0/1.0)
|   |   |   |   TriEdadGenero = Null
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-10s (19.0/9.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   |   |   TriGenero = fem: Female-30s (2.0)
|   |   |   |   |   |   TriGenero = mal
|   |   |   |   |   |   |   LongitudMensaje = Corto: Male-20s (0.0)
|   |   |   |   |   |   |   LongitudMensaje = Medio: Female-20s (2.0/1.0)
|   |   |   |   |   |   |   LongitudMensaje = Largo: Male-20s (6.0/1.0)
|   |   |   |   |   |   TriGenero = sin: Male-30s (3.0/1.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (6.0/2.0)
|   |   |   TriEdad = Alto
|   |   |   |   TriEdadGenero = TF10: Male-30s (0.0)
|   |   |   |   TriEdadGenero = TF20: Male-30s (0.0)
|   |   |   |   TriEdadGenero = TF30: Female-30s (3.0/1.0)
|   |   |   |   TriEdadGenero = TM10: Male-10s (3.0/2.0)
|   |   |   |   TriEdadGenero = TM20: Male-20s (1.0)
|   |   |   |   TriEdadGenero = TM30: Male-30s (2.0)
|   |   |   |   TriEdadGenero = Null
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-30s (8.0/2.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-10s (3.0/2.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-30s (0.0)
|   |   |   TriEdad = Null: Male-20s (15.0/6.0)
|   ErrorEdad = E-20s: Male-20s (3774.0/433.0)
|   ErrorEdad = E-30s
|   |   ErrorAciertosEdadGenero = H-10s
|   |   |   TriEdadGenero = TF10: Male-30s (10.0/3.0)
|   |   |   TriEdadGenero = TF20: Female-20s (2.0)
|   |   |   TriEdadGenero = TF30: Female-30s (51.0/6.0)
|   |   |   TriEdadGenero = TM10: Male-30s (50.0/21.0)
|   |   |   TriEdadGenero = TM20
|   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   TriEdad = Bajo: Male-20s (0.0)
|   |   |   |   |   TriEdad = Medio: Male-20s (2.0)
|   |   |   |   |   TriEdad = Alto: Male-30s (2.0)
|   |   |   |   |   TriEdad = Null: Male-20s (0.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-30s (16.0/4.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Alta
|   |   |   |   |   TriEdad = Bajo: Male-30s (0.0)
|   |   |   |   |   TriEdad = Medio: Male-30s (13.0/5.0)
|   |   |   |   |   TriEdad = Alto
|   |   |   |   |   |   ArgotsInternet = No-Usa: Male-20s (23.0/11.0)
|   |   |   |   |   |   ArgotsInternet = Baja
|   |   |   |   |   |   |   RiquezaVocabulario = Pobre: Male-30s (6.0/2.0)
|   |   |   |   |   |   |   RiquezaVocabulario = Promedio: Female-30s (4.0/2.0)
|   |   |   |   |   |   |   RiquezaVocabulario = Rico: Male-30s (0.0)
|   |   |   |   |   |   ArgotsInternet = Alta: Male-20s (2.0/1.0)
|   |   |   |   |   TriEdad = Null: Male-30s (0.0)
|   |   |   TriEdadGenero = TM30: Male-30s (41.0/3.0)
|   |   |   TriEdadGenero = Null: Male-30s (49.0/19.0)
|   |   ErrorAciertosEdadGenero = H-20s: Male-20s (8.0/2.0)
|   |   ErrorAciertosEdadGenero = H-30s: Male-30s (1571.0/49.0)
|   |   ErrorAciertosEdadGenero = F-10s: Male-30s (121.0/34.0)
|   |   ErrorAciertosEdadGenero = F-20s: Female-20s (3.0/1.0)
|   |   ErrorAciertosEdadGenero = F-30s: Female-30s (52.0/9.0)
|   |   ErrorAciertosEdadGenero = Sin: Male-30s (425.0/86.0)
|   ErrorEdad = Sin
|   |   TriEdadGenero = TF10: Female-30s (2.0)
```

```
|   |   TriEdadGenero = TF20: Male-30s (0.0)
|   |   TriEdadGenero = TF30: Male-20s (31.0/14.0)
|   |   TriEdadGenero = TM10: Male-30s (7.0/3.0)
|   |   TriEdadGenero = TM20
|   |   |   LongitudMensaje = Corto: Male-30s (2.0)
|   |   |   LongitudMensaje = Medio: Male-20s (2.0/1.0)
|   |   |   LongitudMensaje = Largo: Male-20s (1.0)
|   |   TriEdadGenero = TM30: Male-30s (2.0/1.0)
|   |   TriEdadGenero = Null
|   |   |   TriEdad = Bajo: Female-20s (1.0)
|   |   |   TriEdad = Medio: Male-20s (32.0/18.0)
|   |   |   TriEdad = Alto: Male-30s (25.0/9.0)
|   |   |   TriEdad = Null: Male-30s (53.0/30.0)
ErrorGenero = S
|   ErrorEdad = E-10s
|   |   TriEdadGenero = TF10
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   TriEdad = Bajo: Female-30s (4.0/2.0)
|   |   |   |   TriEdad = Medio: Male-10s (3.0/2.0)
|   |   |   |   TriEdad = Alto: Male-10s (0.0)
|   |   |   |   TriEdad = Null: Male-10s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Female-10s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Female-10s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   ArgotsInternet = No-Usa: Female-10s (32.0/17.0)
|   |   |   |   ArgotsInternet = Baja
|   |   |   |   |   RiquezaVocabulario = Pobre: Female-10s (1.0)
|   |   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   |   TriEdad = Bajo: Female-10s (3.0/1.0)
|   |   |   |   |   |   TriEdad = Medio: Male-30s (2.0/1.0)
|   |   |   |   |   |   TriEdad = Alto: Female-10s (0.0)
|   |   |   |   |   |   TriEdad = Null: Female-10s (0.0)
|   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female-20s (3.0/1.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-20s (2.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (0.0)
|   |   |   |   ArgotsInternet = Alta: Female-20s (1.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: Female-10s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: Female-10s (0.0)
|   |   |   ErrorAciertosEdadGenero = Sin: Female-20s (12.0/7.0)
|   |   TriEdadGenero = TF20: Female-20s (8.0/1.0)
|   |   TriEdadGenero = TF30
|   |   |   TriEdad = Bajo
|   |   |   |   RiquezaVocabulario = Pobre: Female-30s (0.0)
|   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   LongitudMensaje = Corto: Female-30s (3.0/1.0)
|   |   |   |   |   LongitudMensaje = Medio: Male-20s (3.0/1.0)
|   |   |   |   |   LongitudMensaje = Largo: Male-20s (0.0)
|   |   |   |   RiquezaVocabulario = Rico: Female-30s (3.0/1.0)
|   |   |   TriEdad = Medio: Female-20s (166.0/100.0)
|   |   |   TriEdad = Alto: Female-30s (11.0/4.0)
|   |   |   TriEdad = Null: Female-20s (0.0)
|   |   TriEdadGenero = TM10
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   LongitudMensaje = Corto: Male-10s (2.0)
|   |   |   |   LongitudMensaje = Medio: Male-10s (2.0/1.0)
|   |   |   |   LongitudMensaje = Largo: Female-20s (9.0/4.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-10s: Female-10s (11.0/6.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   SignosPuntuacion = Usa
```

```
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   |   TriEdad = Bajo: Male-10s (2.0)
|   |   |   |   |   |   TriEdad = Medio: Female-20s (6.0/3.0)
|   |   |   |   |   |   TriEdad = Alto: Male-10s (0.0)
|   |   |   |   |   |   TriEdad = Null: Male-10s (0.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-30s (3.0/1.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (0.0)
|   |   |   |   SignosPuntuacion = No-Usa: Male-20s (2.0)
|   |   TriEdadGenero = TM20: Male-20s (47.0/23.0)
|   |   TriEdadGenero = TM30: Male-30s (8.0)
|   |   TriEdadGenero = Null
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   RiquezaVocabulario = Pobre: Male-10s (5.0/2.0)
|   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   SignosPuntuacion = Usa: Female-20s (5.0/2.0)
|   |   |   |   |   SignosPuntuacion = No-Usa: Male-10s (3.0/1.0)
|   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   |   TriGenero = fem
|   |   |   |   |   |   |   TriEdad = Bajo: Male-20s (2.0)
|   |   |   |   |   |   |   TriEdad = Medio: Male-10s (2.0/1.0)
|   |   |   |   |   |   |   TriEdad = Alto: Male-20s (0.0)
|   |   |   |   |   |   |   TriEdad = Null: Male-20s (0.0)
|   |   |   |   |   |   TriGenero = mal
|   |   |   |   |   |   |   SignosPuntuacion = Usa: Female-30s (17.0/9.0)
|   |   |   |   |   |   |   SignosPuntuacion = No-Usa: Male-20s (4.0/2.0)
|   |   |   |   |   |   TriGenero = sin
|   |   |   |   |   |   |   TriEdad = Bajo
|   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: Male-30s (2.0/1.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Baja: Male-10s (2.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Alta: Male-10s (0.0)
|   |   |   |   |   |   |   TriEdad = Medio: Female-30s (11.0/4.0)
|   |   |   |   |   |   |   TriEdad = Alto: Male-20s (3.0/1.0)
|   |   |   |   |   |   |   TriEdad = Null: Female-30s (21.0/15.0)
|   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   TriGenero = fem: Female-30s (1.0)
|   |   |   |   |   |   TriGenero = mal
|   |   |   |   |   |   |   ArgotsInternet = No-Usa: Female-20s (11.0/6.0)
|   |   |   |   |   |   |   ArgotsInternet = Baja: Male-20s (2.0)
|   |   |   |   |   |   |   ArgotsInternet = Alta: Female-20s (0.0)
|   |   |   |   |   |   TriGenero = sin
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female-30s (4.0/2.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-10s (2.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-10s (0.0)
|   |   |   |   |   LongitudMensaje = Largo: Male-10s (4.0/2.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Male-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Male-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   SignosPuntuacion = Usa
|   |   |   |   |   TriGenero = fem
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   |   |   RiquezaVocabulario = Pobre: Male-20s (3.0/2.0)
|   |   |   |   |   |   |   RiquezaVocabulario = Promedio: Female-10s (3.0/1.0)
|   |   |   |   |   |   |   RiquezaVocabulario = Rico: Female-20s (16.0/10.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-20s (11.0/3.0)
|   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (4.0/2.0)
|   |   |   |   |   TriGenero = mal
|   |   |   |   |   |   RiquezaVocabulario = Pobre: Male-30s (6.0/4.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   |   |   ArgotsInternet = No-Usa: Male-30s (3.0/2.0)
|   |   |   |   |   |   |   ArgotsInternet = Baja: Male-20s (5.0/2.0)
|   |   |   |   |   |   |   ArgotsInternet = Alta: Male-20s (0.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
```

```
|   |   |   |   |   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   |   |   |   |   LongitudMensaje = Corto: Male-30s (18.0/12.0)
|   |   |   |   |   |   |   |   |   LongitudMensaje = Medio: Female-30s (3.0/1.0)
|   |   |   |   |   |   |   |   |   LongitudMensaje = Largo: Male-30s (0.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Baja: Female-20s (4.0/1.0)
|   |   |   |   |   |   |   |   ArgotsInternet = Alta: Male-20s (0.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-30s (8.0/4.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-30s (0.0)
|   |   |   |   |   TriGenero = sin: Male-20s (85.0/56.0)
|   |   |   |   SignosPuntuacion = No-Usa
|   |   |   |   |   ArgotsInternet = No-Usa: Male-30s (40.0/28.0)
|   |   |   |   |   ArgotsInternet = Baja: Male-20s (7.0/3.0)
|   |   |   |   |   ArgotsInternet = Alta: Male-30s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: Male-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: Male-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   RiquezaVocabulario = Pobre: Male-30s (2.0/1.0)
|   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-20s (6.0/2.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-30s (4.0/2.0)
|   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (0.0)
|   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   |   |   TriEdad = Bajo
|   |   |   |   |   |   |   |   TriGenero = fem: Female-20s (1.0)
|   |   |   |   |   |   |   |   TriGenero = mal: Female-20s (2.0)
|   |   |   |   |   |   |   |   TriGenero = sin: Female-30s (4.0/1.0)
|   |   |   |   |   |   |   TriEdad = Medio: Male-20s (23.0/14.0)
|   |   |   |   |   |   |   TriEdad = Alto
|   |   |   |   |   |   |   |   SignosPuntuacion = Usa: Female-30s (5.0/3.0)
|   |   |   |   |   |   |   |   SignosPuntuacion = No-Usa: Male-20s (4.0/1.0)
|   |   |   |   |   |   |   TriEdad = Null
|   |   |   |   |   |   |   |   TriGenero = fem: Male-20s (2.0/1.0)
|   |   |   |   |   |   |   |   TriGenero = mal: Female-20s (5.0/2.0)
|   |   |   |   |   |   |   |   TriGenero = sin: Female-30s (29.0/20.0)
|   |   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   |   TriEdad = Bajo: Female-20s (2.0)
|   |   |   |   |   |   |   TriEdad = Medio
|   |   |   |   |   |   |   |   TriGenero = fem: Female-30s (2.0)
|   |   |   |   |   |   |   |   TriGenero = mal: Male-20s (4.0/2.0)
|   |   |   |   |   |   |   |   TriGenero = sin: Male-30s (4.0/1.0)
|   |   |   |   |   |   |   TriEdad = Alto: Female-30s (3.0/1.0)
|   |   |   |   |   |   |   TriEdad = Null: Female-30s (0.0)
|   |   |   |   |   |   LongitudMensaje = Largo: Male-20s (4.0/2.0)
|   |   |   |   |   ArgotsInternet = Baja
|   |   |   |   |   |   SignosPuntuacion = Usa
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female-20s (4.0/1.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-10s (2.0/1.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (0.0)
|   |   |   |   |   |   SignosPuntuacion = No-Usa: Male-20s (3.0)
|   |   |   |   |   ArgotsInternet = Alta: Male-20s (0.0)
|   ErrorEdad = E-20s
|   |   TriEdadGenero = TF10: Female-20s (165.0/85.0)
|   |   TriEdadGenero = TF20: Female-20s (44.0/2.0)
|   |   TriEdadGenero = TF30
|   |   |   TriEdad = Bajo
|   |   |   |   ErrorAciertosEdadGenero = H-10s: Male-30s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = H-20s: Female-20s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = H-30s: Male-20s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-10s: Male-20s (6.0/2.0)
|   |   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = F-30s: Female-30s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = Sin: Female-20s (14.0/9.0)
```

```
|   |   |   TriEdad = Medio
|   |   |   |   ErrorAciertosEdadGenero = H-10s: Male-20s (30.0/14.0)
|   |   |   |   ErrorAciertosEdadGenero = H-20s: Male-20s (3.0)
|   |   |   |   ErrorAciertosEdadGenero = H-30s
|   |   |   |   |   ArgotsInternet = No-Usa: Male-30s (13.0/6.0)
|   |   |   |   |   ArgotsInternet = Baja: Female-20s (2.0/1.0)
|   |   |   |   |   ArgotsInternet = Alta: Male-30s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-10s: Female-20s (133.0/64.0)
|   |   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (27.0/5.0)
|   |   |   |   ErrorAciertosEdadGenero = F-30s: Female-20s (14.0/8.0)
|   |   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   |   TriGenero = fem
|   |   |   |   |   |   LongitudMensaje = Corto: Female-20s (193.0/124.0)
|   |   |   |   |   |   LongitudMensaje = Medio: Female-20s (84.0/39.0)
|   |   |   |   |   |   LongitudMensaje = Largo: Male-20s (22.0/9.0)
|   |   |   |   |   TriGenero = mal
|   |   |   |   |   |   ArgotsInternet = No-Usa: Female-20s (12.0/6.0)
|   |   |   |   |   |   ArgotsInternet = Baja: Male-20s (2.0)
|   |   |   |   |   |   ArgotsInternet = Alta: Female-20s (0.0)
|   |   |   |   |   TriGenero = sin
|   |   |   |   |   |   LongitudMensaje = Corto: Female-20s (11.0/5.0)
|   |   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-20s (6.0/3.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-30s (4.0/1.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (0.0)
|   |   |   |   |   |   LongitudMensaje = Largo: Female-20s (4.0/1.0)
|   |   |   TriEdad = Alto: Female-30s (34.0/16.0)
|   |   |   TriEdad = Null: Female-20s (0.0)
|   |   TriEdadGenero = TM10
|   |   |   TriEdad = Bajo
|   |   |   |   TriGenero = fem: Female-20s (2.0)
|   |   |   |   TriGenero = mal
|   |   |   |   |   RiquezaVocabulario = Pobre: Male-10s (3.0/1.0)
|   |   |   |   |   RiquezaVocabulario = Promedio: Male-30s (5.0/3.0)
|   |   |   |   |   RiquezaVocabulario = Rico: Male-20s (26.0/15.0)
|   |   |   |   TriGenero = sin: Female-30s (1.0)
|   |   |   TriEdad = Medio
|   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-20s (19.0/8.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   |   RiquezaVocabulario = Pobre: Male-20s (2.0/1.0)
|   |   |   |   |   |   RiquezaVocabulario = Promedio: Female-20s (6.0/2.0)
|   |   |   |   |   |   RiquezaVocabulario = Rico: Male-20s (3.0/1.0)
|   |   |   |   |   ArgotsInternet = Baja: Male-20s (12.0/4.0)
|   |   |   |   |   ArgotsInternet = Alta: Male-20s (0.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (12.0/4.0)
|   |   |   TriEdad = Alto: Female-20s (14.0/7.0)
|   |   |   TriEdad = Null: Male-20s (0.0)
|   |   TriEdadGenero = TM20
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-20s (14.0/4.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-20s (4.0/1.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (4.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Male-20s (3.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Male-20s (2.0/1.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   SignosPuntuacion = Usa: Male-20s (67.0/20.0)
|   |   |   |   SignosPuntuacion = No-Usa: Female-20s (3.0/1.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (12.0/2.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: Male-20s (7.0/2.0)
|   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   TriGenero = fem: Male-20s (0.0)
|   |   |   |   TriGenero = mal: Male-20s (115.0/38.0)
|   |   |   |   TriGenero = sin: Female-20s (4.0)
```

```
|   |   TriEdadGenero = TM30: Male-30s (23.0/7.0)
|   |   TriEdadGenero = Null
|   |   |   ErrorAciertosEdadGenero = H-10s: Male-20s (75.0/45.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Male-20s (17.0/5.0)
|   |   |   ErrorAciertosEdadGenero = H-30s
|   |   |   |   RiquezaVocabulario = Pobre: Male-20s (3.0/2.0)
|   |   |   |   RiquezaVocabulario = Promedio: Male-20s (2.0)
|   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   ArgotsInternet = No-Usa: Male-30s (27.0/12.0)
|   |   |   |   |   ArgotsInternet = Baja: Male-20s (3.0)
|   |   |   |   |   ArgotsInternet = Alta: Male-30s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   TriGenero = fem
|   |   |   |   |   TriEdad = Bajo: Male-20s (2.0)
|   |   |   |   |   TriEdad = Medio
|   |   |   |   |   |   ArgotsInternet = No-Usa: Female-20s (33.0/17.0)
|   |   |   |   |   |   ArgotsInternet = Baja: Male-20s (8.0/3.0)
|   |   |   |   |   |   ArgotsInternet = Alta: Female-20s (0.0)
|   |   |   |   |   TriEdad = Alto: Female-30s (3.0/1.0)
|   |   |   |   |   TriEdad = Null: Male-30s (4.0/2.0)
|   |   |   |   TriGenero = mal: Male-20s (114.0/55.0)
|   |   |   |   TriGenero = sin
|   |   |   |   |   RiquezaVocabulario = Pobre: Female-20s (2.0/1.0)
|   |   |   |   |   RiquezaVocabulario = Promedio: Male-20s (18.0/8.0)
|   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   TriEdad = Bajo: Male-20s (7.0/2.0)
|   |   |   |   |   |   TriEdad = Medio
|   |   |   |   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   |   |   |   |   |   SignosPuntuacion = Usa
|   |   |   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: Male-30s (31.0/21.0)
|   |   |   |   |   |   |   |   |   |   ArgotsInternet = Baja: Female-20s (2.0/1.0)
|   |   |   |   |   |   |   |   |   |   ArgotsInternet = Alta: Male-30s (0.0)
|   |   |   |   |   |   |   |   |   SignosPuntuacion = No-Usa
|   |   |   |   |   |   |   |   |   |   ArgotsInternet = No-Usa: Female-20s (15.0/8.0)
|   |   |   |   |   |   |   |   |   |   ArgotsInternet = Baja: Male-20s (4.0/3.0)
|   |   |   |   |   |   |   |   |   |   ArgotsInternet = Alta: Female-20s (0.0)
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-20s (2.0)
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (0.0)
|   |   |   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-20s (3.0)
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-30s (17.0/11.0)
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (0.0)
|   |   |   |   |   |   |   LongitudMensaje = Largo: Male-20s (4.0/2.0)
|   |   |   |   |   |   TriEdad = Alto
|   |   |   |   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   |   |   |   SignosPuntuacion = Usa: Female-30s (7.0/2.0)
|   |   |   |   |   |   |   |   SignosPuntuacion = No-Usa: Male-20s (2.0/1.0)
|   |   |   |   |   |   |   LongitudMensaje = Medio: Female-20s (3.0/1.0)
|   |   |   |   |   |   |   LongitudMensaje = Largo: Female-30s (0.0)
|   |   |   |   |   |   TriEdad = Null
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-20s (70.0/50.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-20s (4.0/1.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (58.0/16.0)
|   |   |   ErrorAciertosEdadGenero = F-30s
|   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   TriGenero = fem: Female-20s (1.0)
|   |   |   |   |   TriGenero = mal: Female-30s (10.0/1.0)
|   |   |   |   |   TriGenero = sin: Male-20s (22.0/13.0)
|   |   |   |   LongitudMensaje = Medio: Male-20s (8.0/4.0)
|   |   |   |   LongitudMensaje = Largo: Male-20s (3.0/1.0)
|   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   RiquezaVocabulario = Pobre: Male-20s (37.0/22.0)
```

```
|   |   |   |   RiquezaVocabulario = Promedio
|   |   |   |   |   TriGenero = fem: Female-20s (26.0/13.0)
|   |   |   |   |   TriGenero = mal: Male-20s (32.0/15.0)
|   |   |   |   |   TriGenero = sin: Female-20s (38.0/23.0)
|   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   TriEdad = Bajo
|   |   |   |   |   |   SignosPuntuacion = Usa: Male-20s (17.0/11.0)
|   |   |   |   |   |   SignosPuntuacion = No-Usa
|   |   |   |   |   |   |   TriGenero = fem: Female-20s (0.0)
|   |   |   |   |   |   |   TriGenero = mal: Male-10s (2.0/1.0)
|   |   |   |   |   |   |   TriGenero = sin: Female-20s (8.0/4.0)
|   |   |   |   |   TriEdad = Medio
|   |   |   |   |   |   TriGenero = fem
|   |   |   |   |   |   |   SignosPuntuacion = Usa: Female-20s (47.0/24.0)
|   |   |   |   |   |   |   SignosPuntuacion = No-Usa: Male-30s (2.0/1.0)
|   |   |   |   |   |   TriGenero = mal
|   |   |   |   |   |   |   LongitudMensaje = Corto: Male-20s (59.0/33.0)
|   |   |   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-20s (10.0/3.0)
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-30s (28.0/18.0)
|   |   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (0.0)
|   |   |   |   |   |   |   LongitudMensaje = Largo: Female-20s (8.0/2.0)
|   |   |   |   |   |   TriGenero = sin: Male-20s (218.0/145.0)
|   |   |   |   |   TriEdad = Alto
|   |   |   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-30s (79.0/46.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-20s (26.0/16.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-30s (0.0)
|   |   |   |   |   |   ArgotsInternet = Baja
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Baja: Male-20s (2.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Media: Female-20s (2.0)
|   |   |   |   |   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (0.0)
|   |   |   |   |   |   ArgotsInternet = Alta: Male-30s (0.0)
|   |   |   |   |   TriEdad = Null
|   |   |   |   |   |   UsoPalabrasCerradas = Usa: Male-20s (364.0/237.0)
|   |   |   |   |   |   UsoPalabrasCerradas = No-Usa: Male-30s (6.0/2.0)
|   ErrorEdad = E-30s
|   |   ErrorAciertosEdadGenero = H-10s: Female-30s (70.0/37.0)
|   |   ErrorAciertosEdadGenero = H-20s: Female-30s (0.0)
|   |   ErrorAciertosEdadGenero = H-30s: Male-30s (114.0/23.0)
|   |   ErrorAciertosEdadGenero = F-10s
|   |   |   TriEdad = Bajo: Male-30s (8.0/3.0)
|   |   |   TriEdad = Medio
|   |   |   |   TriGenero = fem: Female-30s (21.0/5.0)
|   |   |   |   TriGenero = mal
|   |   |   |   |   RiquezaVocabulario = Pobre: Male-20s (4.0/3.0)
|   |   |   |   |   RiquezaVocabulario = Promedio: Male-30s (9.0/3.0)
|   |   |   |   |   RiquezaVocabulario = Rico: Female-30s (11.0/6.0)
|   |   |   |   TriGenero = sin
|   |   |   |   |   ArgotsInternet = No-Usa: Female-30s (12.0/4.0)
|   |   |   |   |   ArgotsInternet = Baja: Male-30s (2.0/1.0)
|   |   |   |   |   ArgotsInternet = Alta: Female-30s (0.0)
|   |   |   TriEdad = Alto
|   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   TriEdadGenero = TF10: Male-30s (0.0)
|   |   |   |   |   TriEdadGenero = TF20: Male-30s (0.0)
|   |   |   |   |   TriEdadGenero = TF30: Female-30s (4.0/1.0)
|   |   |   |   |   TriEdadGenero = TM10: Male-30s (1.0)
|   |   |   |   |   TriEdadGenero = TM20: Female-30s (2.0)
|   |   |   |   |   TriEdadGenero = TM30: Male-30s (3.0/1.0)
|   |   |   |   |   TriEdadGenero = Null: Male-30s (15.0/7.0)
|   |   |   |   ArgotsInternet = Baja: Male-30s (3.0)
|   |   |   |   ArgotsInternet = Alta: Female-30s (1.0)
|   |   |   TriEdad = Null: Male-30s (12.0/4.0)
```

```
|   |   ErrorAciertosEdadGenero = F-20s: Female-20s (1.0)
|   |   ErrorAciertosEdadGenero = F-30s: Female-30s (126.0/25.0)
|   |   ErrorAciertosEdadGenero = Sin
|   |   |   TriEdadGenero = TF10
|   |   |   |   FrecuenciaPalabrasCerradas = Baja: Female-30s (16.0/4.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Media: Male-30s (5.0/2.0)
|   |   |   |   FrecuenciaPalabrasCerradas = Alta: Female-30s (0.0)
|   |   |   TriEdadGenero = TF20: Female-20s (3.0)
|   |   |   TriEdadGenero = TF30: Female-30s (63.0/30.0)
|   |   |   TriEdadGenero = TM10
|   |   |   |   ArgotsInternet = No-Usa: Male-30s (15.0/6.0)
|   |   |   |   ArgotsInternet = Baja: Female-20s (2.0/1.0)
|   |   |   |   ArgotsInternet = Alta: Female-30s (1.0)
|   |   |   TriEdadGenero = TM20: Female-30s (15.0/5.0)
|   |   |   TriEdadGenero = TM30: Male-30s (13.0/3.0)
|   |   |   TriEdadGenero = Null
|   |   |   |   RiquezaVocabulario = Pobre
|   |   |   |   |   TriGenero = fem: Female-20s (2.0)
|   |   |   |   |   TriGenero = mal: Male-30s (2.0/1.0)
|   |   |   |   |   TriGenero = sin: Female-20s (0.0)
|   |   |   |   RiquezaVocabulario = Promedio: Female-30s (26.0/14.0)
|   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   |   TriEdad = Bajo: Female-30s (5.0/3.0)
|   |   |   |   |   |   TriEdad = Medio: Female-30s (42.0/19.0)
|   |   |   |   |   |   TriEdad = Alto: Male-30s (49.0/25.0)
|   |   |   |   |   |   TriEdad = Null: Male-30s (71.0/36.0)
|   |   |   |   |   LongitudMensaje = Medio
|   |   |   |   |   |   TriEdad = Bajo: Female-30s (1.0)
|   |   |   |   |   |   TriEdad = Medio: Male-30s (20.0/8.0)
|   |   |   |   |   |   TriEdad = Alto: Female-30s (31.0/13.0)
|   |   |   |   |   |   TriEdad = Null: Female-30s (7.0/3.0)
|   |   |   |   |   LongitudMensaje = Largo
|   |   |   |   |   |   TriGenero = fem: Male-30s (2.0)
|   |   |   |   |   |   TriGenero = mal: Male-30s (5.0/1.0)
|   |   |   |   |   |   TriGenero = sin: Female-30s (7.0/1.0)
|   ErrorEdad = Sin
|   |   TriEdadGenero = TF10
|   |   |   ErrorAciertosEdadGenero = H-10s: Female-30s (3.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   RiquezaVocabulario = Pobre: Female-20s (1.0)
|   |   |   |   RiquezaVocabulario = Promedio: Female-20s (2.0)
|   |   |   |   RiquezaVocabulario = Rico: Female-30s (4.0/2.0)
|   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: Male-20s (1.0)
|   |   |   ErrorAciertosEdadGenero = Sin: Female-20s (57.0/36.0)
|   |   TriEdadGenero = TF20: Female-20s (12.0)
|   |   TriEdadGenero = TF30
|   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   LongitudMensaje = Corto
|   |   |   |   |   RiquezaVocabulario = Pobre: Male-20s (1.0)
|   |   |   |   |   RiquezaVocabulario = Promedio: Female-20s (5.0/2.0)
|   |   |   |   |   RiquezaVocabulario = Rico: Female-30s (7.0/2.0)
|   |   |   |   LongitudMensaje = Medio: Female-20s (2.0)
|   |   |   |   LongitudMensaje = Largo: Male-30s (1.0)
|   |   |   ErrorAciertosEdadGenero = H-20s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = H-30s: Male-30s (2.0/1.0)
|   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   TriEdad = Bajo: Male-20s (3.0/2.0)
|   |   |   |   TriEdad = Medio: Female-20s (31.0/15.0)
|   |   |   |   TriEdad = Alto: Female-30s (3.0)
|   |   |   |   TriEdad = Null: Female-20s (0.0)
```

```
|   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (0.0)
|   |   |   ErrorAciertosEdadGenero = F-30s: Female-30s (7.0/2.0)
|   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   TriEdad = Bajo: Female-30s (9.0/5.0)
|   |   |   |   |   TriEdad = Medio: Female-20s (217.0/142.0)
|   |   |   |   |   TriEdad = Alto: Female-30s (22.0/12.0)
|   |   |   |   |   TriEdad = Null: Female-20s (0.0)
|   |   |   |   ArgotsInternet = Baja: Male-20s (10.0/6.0)
|   |   |   |   ArgotsInternet = Alta: Female-20s (0.0)
|   |   TriEdadGenero = TM10
|   |   |   TriGenero = fem: Male-20s (2.0/1.0)
|   |   |   TriGenero = mal
|   |   |   |   ErrorAciertosEdadGenero = H-10s: Female-20s (3.0/1.0)
|   |   |   |   ErrorAciertosEdadGenero = H-20s: Male-20s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = H-30s: Male-20s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-10s: Male-20s (2.0)
|   |   |   |   ErrorAciertosEdadGenero = F-20s: Male-20s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-30s: Female-20s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = Sin: Male-30s (20.0/13.0)
|   |   |   TriGenero = sin: Female-20s (1.0)
|   |   TriEdadGenero = TM20: Male-20s (42.0/19.0)
|   |   TriEdadGenero = TM30
|   |   |   TriGenero = fem: Male-30s (0.0)
|   |   |   TriGenero = mal: Male-30s (13.0)
|   |   |   TriGenero = sin: Female-30s (2.0)
|   |   TriEdadGenero = Null
|   |   |   FrecuenciaPalabrasCerradas = Baja
|   |   |   |   ErrorAciertosEdadGenero = H-10s
|   |   |   |   |   TriEdad = Bajo: Male-20s (2.0/1.0)
|   |   |   |   |   TriEdad = Medio: Male-20s (17.0/11.0)
|   |   |   |   |   TriEdad = Alto: Female-20s (4.0/2.0)
|   |   |   |   |   TriEdad = Null
|   |   |   |   |   |   SignosPuntuacion = Usa: Female-30s (7.0/3.0)
|   |   |   |   |   |   SignosPuntuacion = No-Usa: Male-10s (3.0/2.0)
|   |   |   |   ErrorAciertosEdadGenero = H-20s: Male-30s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = H-30s: Male-30s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   |   RiquezaVocabulario = Pobre: Male-30s (2.0/1.0)
|   |   |   |   |   RiquezaVocabulario = Promedio: Female-20s (2.0)
|   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   TriGenero = fem: Male-30s (12.0/6.0)
|   |   |   |   |   |   TriGenero = mal
|   |   |   |   |   |   |   TriEdad = Bajo: Male-30s (0.0)
|   |   |   |   |   |   |   TriEdad = Medio: Male-30s (5.0)
|   |   |   |   |   |   |   TriEdad = Alto: Female-30s (4.0/1.0)
|   |   |   |   |   |   |   TriEdad = Null: Female-30s (3.0/1.0)
|   |   |   |   |   |   TriGenero = sin
|   |   |   |   |   |   |   TriEdad = Bajo: Male-30s (0.0)
|   |   |   |   |   |   |   TriEdad = Medio: Female-30s (21.0/14.0)
|   |   |   |   |   |   |   TriEdad = Alto: Male-30s (9.0/4.0)
|   |   |   |   |   |   |   TriEdad = Null: Male-30s (41.0/26.0)
|   |   |   |   ErrorAciertosEdadGenero = F-20s: Female-30s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-30s
|   |   |   |   |   SignosPuntuacion = Usa
|   |   |   |   |   |   TriEdad = Bajo: Female-30s (0.0)
|   |   |   |   |   |   TriEdad = Medio: Male-20s (4.0/2.0)
|   |   |   |   |   |   TriEdad = Alto: Female-30s (0.0)
|   |   |   |   |   |   TriEdad = Null: Female-30s (3.0)
|   |   |   |   |   SignosPuntuacion = No-Usa: Male-30s (2.0/1.0)
|   |   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   |   ArgotsInternet = No-Usa
|   |   |   |   |   |   SignosPuntuacion = Usa: Female-30s (621.0/418.0)
|   |   |   |   |   |   SignosPuntuacion = No-Usa: Male-30s (324.0/220.0)
```

```
|   |   |   |   |   ArgotsInternet = Baja: Female-20s (37.0/24.0)
|   |   |   |   |   ArgotsInternet = Alta: Female-30s (0.0)
|   |   |   FrecuenciaPalabrasCerradas = Media
|   |   |   |   ErrorAciertosEdadGenero = H-10s: Female-20s (6.0/2.0)
|   |   |   |   ErrorAciertosEdadGenero = H-20s: Female-20s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = H-30s: Male-30s (1.0)
|   |   |   |   ErrorAciertosEdadGenero = F-10s
|   |   |   |   |   TriGenero = fem: Male-20s (4.0/1.0)
|   |   |   |   |   TriGenero = mal
|   |   |   |   |   |   LongitudMensaje = Corto: Male-20s (0.0)
|   |   |   |   |   |   LongitudMensaje = Medio: Male-20s (3.0/1.0)
|   |   |   |   |   |   LongitudMensaje = Largo: Female-20s (2.0/1.0)
|   |   |   |   |   TriGenero = sin: Male-30s (9.0/4.0)
|   |   |   |   ErrorAciertosEdadGenero = F-20s: Female-20s (0.0)
|   |   |   |   ErrorAciertosEdadGenero = F-30s: Male-30s (2.0)
|   |   |   |   ErrorAciertosEdadGenero = Sin
|   |   |   |   |   RiquezaVocabulario = Pobre: Female-20s (5.0/2.0)
|   |   |   |   |   RiquezaVocabulario = Promedio: Female-20s (10.0/5.0)
|   |   |   |   |   RiquezaVocabulario = Rico
|   |   |   |   |   |   TriEdad = Bajo: Female-20s (2.0)
|   |   |   |   |   |   TriEdad = Medio: Male-20s (25.0/16.0)
|   |   |   |   |   |   TriEdad = Alto: Male-20s (8.0/5.0)
|   |   |   |   |   |   TriEdad = Null: Female-20s (10.0/3.0)
|   |   |   FrecuenciaPalabrasCerradas = Alta: Male-20s (2.0/1.0)

Number of Leaves  :       576

Size of the tree :        788


Time taken to build model: 0.13 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13860               69.3    %
Incorrectly Classified Instances      6140               30.7    %
Kappa statistic                          0.5939
Mean absolute error                      0.1302
Root mean squared error                  0.2634
Relative absolute error                 51.4392 %
Root relative squared error             74.0583 %
Total Number of Instances            20000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.584     0.007      0.574     0.584     0.579      0.906    Male-10s
                 0.743     0.125      0.703     0.743     0.722      0.887    Male-20s
                 0.641     0.063      0.716     0.641     0.677      0.872    Male-30s
                 0.482     0.006      0.605     0.482     0.536      0.884    Female-10s
                 0.736     0.122      0.704     0.736     0.719      0.882    Female-20s
                 0.641     0.084      0.658     0.641     0.649      0.872    Female-30s
Weighted Avg.    0.693     0.1        0.693     0.693     0.692      0.88

=== Confusion Matrix ===

    a    b    c    d    e    f   <-- classified as
  194   60   28    0   29   21 |    a = Male-10s
   61 4236  378   16  700  314 |    b = Male-20s
   40  626 2542    6  276  473 |    c = Male-30s
    1   37   13  170   99   33 |    d = Female-10s
   27  713  199   57 4152  495 |    e = Female-20s
   15  356  392   32  643 2566 |    f = Female-30s
```

Continuamos con el algoritmo Naive Bayes y a continuación mostramos los resultados obtenidos.

```
=== Run information ===

Scheme:weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -
P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R14-15
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
              ErrorAciertosEdadGenero
              General
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Bayes Network Classifier
not using ADTree
#attributes=14 #classindex=13
Network structure (nodes followed by parents)
UsoPalabrasCerradas(2): General
FrecuenciaPalabrasCerradas(3): General
RiquezaVocabulario(3): General
SignosPuntuacion(2): General
LongitudMensaje(3): General
ArgotsInternet(3): General
Emoticones(3): General
TriEdad(4): General
TriEdadGenero(7): General
TriGenero(3): General
ErrorGenero(3): General
ErrorEdad(4): General
ErrorAciertosEdadGenero(7): General
General(6):
LogScore Bayes: -242481.66952624178
LogScore BDeu: -243086.636434834
LogScore MDL: -243079.56844286376
LogScore ENTROPY: -242044.65399362374
LogScore AIC: -242253.65399362374


Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13868               69.34   %
Incorrectly Classified Instances      6132               30.66   %
Kappa statistic                          0.5966
Mean absolute error                      0.1213
Root mean squared error                  0.263
Relative absolute error                 47.944  %
```

```
Root relative squared error            73.9509 %
Total Number of Instances           20000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.581     0.011      0.475     0.581     0.523      0.967    Male-10s
                 0.745     0.128      0.698     0.745     0.721      0.904    Male-20s
                 0.661     0.067      0.708     0.661     0.683      0.901    Male-30s
                 0.637     0.013      0.476     0.637     0.545      0.967    Female-10s
                 0.718     0.095      0.747     0.718     0.732      0.902    Female-20s
                 0.632     0.086      0.647     0.632     0.64       0.898    Female-30s
Weighted Avg.    0.693     0.095      0.696     0.693     0.694      0.904

=== Confusion Matrix ===

    a    b    c    d    e    f   <-- classified as
  193   89   18    3   19   10 |    a = Male-10s
   73 4251  414   47  592  328 |    b = Male-20s
   49  626 2618   13  171  486 |    c = Male-30s
    4   31    5  225   70   18 |    d = Female-10s
   45  699  183  127 4049  540 |    e = Female-20s
   42  391  462   58  519 2532 |    f = Female-30s
```

Por ultimo para terminar esta prueba, le aplicamos el algoritmo Maquinas de Soporte Vectorial y a continuación mostramos los resultados obtenidos.

```
=== Run information ===

Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K
"weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
Relation:     Conversaciones-weka.filters.unsupervised.attribute.Remove-R14-15
Instances:    20000
Attributes:   14
              UsoPalabrasCerradas
              FrecuenciaPalabrasCerradas
              RiquezaVocabulario
              SignosPuntuacion
              LongitudMensaje
              ArgotsInternet
              Emoticones
              TriEdad
              TriEdadGenero
              TriGenero
              ErrorGenero
              ErrorEdad
              ErrorAciertosEdadGenero
              General
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: Male-10s, Male-20s

BinarySMO
Machine linear: showing attribute weights, not support vectors.

        -0.3032 * (normalized) UsoPalabrasCerradas
 +       0.0334 * (normalized) FrecuenciaPalabrasCerradas=Baja
```

```
 +       0.0328 * (normalized) FrecuenciaPalabrasCerradas=Media
 +      -0.0662 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.3098 * (normalized) RiquezaVocabulario=Pobre
 +      -0.0774 * (normalized) RiquezaVocabulario=Promedio
 +       0.3873 * (normalized) RiquezaVocabulario=Rico
 +      -0.1651 * (normalized) SignosPuntuacion
 +      -0.3664 * (normalized) LongitudMensaje=Corto
 +      -0.0673 * (normalized) LongitudMensaje=Medio
 +       0.4337 * (normalized) LongitudMensaje=Largo
 +       0.1102 * (normalized) ArgotsInternet=No-Usa
 +      -0.0557 * (normalized) ArgotsInternet=Baja
 +      -0.0545 * (normalized) ArgotsInternet=Alta
 +       0.7355 * (normalized) Emoticones=No-Usa
 +       0.1658 * (normalized) Emoticones=Baja
 +      -0.9013 * (normalized) Emoticones=Alta
 +      -0.534  * (normalized) TriEdad=Bajo
 +       0.1003 * (normalized) TriEdad=Medio
 +       0.3337 * (normalized) TriEdad=Alto
 +       0.1    * (normalized) TriEdad=Null
 +      -0.5212 * (normalized) TriEdadGenero=TF10
 +       0.1127 * (normalized) TriEdadGenero=TF30
 +      -0.9548 * (normalized) TriEdadGenero=TM10
 +       0.6763 * (normalized) TriEdadGenero=TM20
 +       0.9088 * (normalized) TriEdadGenero=TM30
 +      -0.2218 * (normalized) TriEdadGenero=Null
 +       0.0772 * (normalized) TriGenero=fem
 +       0.1787 * (normalized) TriGenero=mal
 +      -0.2559 * (normalized) TriGenero=sin
 +       0.4094 * (normalized) ErrorGenero=F
 +      -0.4227 * (normalized) ErrorGenero=M
 +       0.0133 * (normalized) ErrorGenero=S
 +      -1.0671 * (normalized) ErrorEdad=E-10s
 +       0.8012 * (normalized) ErrorEdad=E-20s
 +       0.3975 * (normalized) ErrorEdad=E-30s
 +      -0.1316 * (normalized) ErrorEdad=Sin
 +      -1.1272 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0.7351 * (normalized) ErrorAciertosEdadGenero=H-20s
 +       0.2429 * (normalized) ErrorAciertosEdadGenero=H-30s
 +       0.6431 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.4939 * (normalized) ErrorAciertosEdadGenero=Sin
 +       1.6915

Number of kernel evaluations: 6221213 (61.478% cached)

Classifier for classes: Male-10s, Male-30s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

        -0.1004 * (normalized) UsoPalabrasCerradas
 +       0.09   * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.098  * (normalized) FrecuenciaPalabrasCerradas=Media
 +      -0.188  * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.2876 * (normalized) RiquezaVocabulario=Pobre
 +      -0.1662 * (normalized) RiquezaVocabulario=Promedio
 +       0.4539 * (normalized) RiquezaVocabulario=Rico
 +      -0.5972 * (normalized) SignosPuntuacion
 +      -0.4507 * (normalized) LongitudMensaje=Corto
 +      -0.146  * (normalized) LongitudMensaje=Medio
 +       0.5967 * (normalized) LongitudMensaje=Largo
 +       0.1134 * (normalized) ArgotsInternet=No-Usa
 +      -0.1992 * (normalized) ArgotsInternet=Baja
 +       0.0857 * (normalized) ArgotsInternet=Alta
```

```
 +       1.138  * (normalized) Emoticones=No-Usa
 +      -0.9132 * (normalized) Emoticones=Baja
 +      -0.2248 * (normalized) Emoticones=Alta
 +      -0.4473 * (normalized) TriEdad=Bajo
 +      -0.164  * (normalized) TriEdad=Medio
 +       0.5525 * (normalized) TriEdad=Alto
 +       0.0588 * (normalized) TriEdad=Null
 +      -0.227  * (normalized) TriEdadGenero=TF10
 +       0.0967 * (normalized) TriEdadGenero=TF30
 +      -1.0473 * (normalized) TriEdadGenero=TM10
 +       0.1472 * (normalized) TriEdadGenero=TM20
 +       1.1439 * (normalized) TriEdadGenero=TM30
 +      -0.1135 * (normalized) TriEdadGenero=Null
 +      -0.2099 * (normalized) TriGenero=fem
 +       0.2732 * (normalized) TriGenero=mal
 +      -0.0633 * (normalized) TriGenero=sin
 +       0.433  * (normalized) ErrorGenero=F
 +      -0.4182 * (normalized) ErrorGenero=M
 +      -0.0148 * (normalized) ErrorGenero=S
 +      -1.4215 * (normalized) ErrorEdad=E-10s
 +       0.1748 * (normalized) ErrorEdad=E-20s
 +       1.1747 * (normalized) ErrorEdad=E-30s
 +       0.072  * (normalized) ErrorEdad=Sin
 +      -1.1712 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-20s
 +       1.0073 * (normalized) ErrorAciertosEdadGenero=H-30s
 +       0.7385 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.5746 * (normalized) ErrorAciertosEdadGenero=Sin
 +       1.2912

Number of kernel evaluations: 2402593 (73.184% cached)

Classifier for classes: Male-10s, Female-10s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

         0      * (normalized) UsoPalabrasCerradas
 +      -0.0019 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +      -0.0121 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.014  * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.0089 * (normalized) RiquezaVocabulario=Pobre
 +      -0.0075 * (normalized) RiquezaVocabulario=Promedio
 +       0.0164 * (normalized) RiquezaVocabulario=Rico
 +      -0.023  * (normalized) SignosPuntuacion
 +      -0.0168 * (normalized) LongitudMensaje=Corto
 +      -0.0077 * (normalized) LongitudMensaje=Medio
 +       0.0245 * (normalized) LongitudMensaje=Largo
 +       0.005  * (normalized) ArgotsInternet=No-Usa
 +      -0.005  * (normalized) ArgotsInternet=Baja
 +       0      * (normalized) ArgotsInternet=Alta
 +       0.021  * (normalized) Emoticones=No-Usa
 +      -0.0512 * (normalized) Emoticones=Baja
 +       0.0302 * (normalized) Emoticones=Alta
 +      -0.015  * (normalized) TriEdad=Bajo
 +      -0.0175 * (normalized) TriEdad=Medio
 +       0.0276 * (normalized) TriEdad=Alto
 +       0.0049 * (normalized) TriEdad=Null
 +       0.3788 * (normalized) TriEdadGenero=TF10
 +      -0.1087 * (normalized) TriEdadGenero=TF30
 +      -0.076  * (normalized) TriEdadGenero=TM10
 +      -0.0741 * (normalized) TriEdadGenero=TM20
 +      -0.1199 * (normalized) TriEdadGenero=Null
```

```
 +       0.0188 * (normalized) TriGenero=fem
 +      -0.0283 * (normalized) TriGenero=mal
 +       0.0095 * (normalized) TriGenero=sin
 +       0.7445 * (normalized) ErrorGenero=F
 +      -0.6212 * (normalized) ErrorGenero=M
 +      -0.1233 * (normalized) ErrorGenero=S
 +      -0.0044 * (normalized) ErrorEdad=E-10s
 +       0.0122 * (normalized) ErrorEdad=E-20s
 +      -0.0078 * (normalized) ErrorEdad=Sin
 +      -1.2555 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       1.6334 * (normalized) ErrorAciertosEdadGenero=F-10s
 +      -0.3778 * (normalized) ErrorAciertosEdadGenero=Sin
 -       0.386

Number of kernel evaluations: 104176 (87.77% cached)

Classifier for classes: Male-10s, Female-20s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

        -0.2423 * (normalized) UsoPalabrasCerradas
 +      -0.0642 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.0035 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.0607 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.2455 * (normalized) RiquezaVocabulario=Pobre
 +      -0.188  * (normalized) RiquezaVocabulario=Promedio
 +       0.4335 * (normalized) RiquezaVocabulario=Rico
 +      -0.6941 * (normalized) SignosPuntuacion
 +      -0.3847 * (normalized) LongitudMensaje=Corto
 +      -0.2269 * (normalized) LongitudMensaje=Medio
 +       0.6116 * (normalized) LongitudMensaje=Largo
 +       0.0188 * (normalized) ArgotsInternet=No-Usa
 +      -0.0381 * (normalized) ArgotsInternet=Baja
 +       0.0192 * (normalized) ArgotsInternet=Alta
 +       0.502  * (normalized) Emoticones=No-Usa
 +       0.0021 * (normalized) Emoticones=Baja
 +      -0.5042 * (normalized) Emoticones=Alta
 +      -0.6349 * (normalized) TriEdad=Bajo
 +       0.0462 * (normalized) TriEdad=Medio
 +       0.4164 * (normalized) TriEdad=Alto
 +       0.1723 * (normalized) TriEdad=Null
 +      -0.255  * (normalized) TriEdadGenero=TF10
 +       0.7781 * (normalized) TriEdadGenero=TF20
 +      -0.0018 * (normalized) TriEdadGenero=TF30
 +      -0.9083 * (normalized) TriEdadGenero=TM10
 +      -0.0849 * (normalized) TriEdadGenero=TM20
 +       0.9384 * (normalized) TriEdadGenero=TM30
 +      -0.4665 * (normalized) TriEdadGenero=Null
 +       0.1128 * (normalized) TriGenero=fem
 +       0.0561 * (normalized) TriGenero=mal
 +      -0.1688 * (normalized) TriGenero=sin
 +       0.865  * (normalized) ErrorGenero=F
 +      -1.1349 * (normalized) ErrorGenero=M
 +       0.2699 * (normalized) ErrorGenero=S
 +      -1.0307 * (normalized) ErrorEdad=E-10s
 +       0.4699 * (normalized) ErrorEdad=E-20s
 +       0.6559 * (normalized) ErrorEdad=E-30s
 +      -0.0951 * (normalized) ErrorEdad=Sin
 +      -1.1267 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0.2926 * (normalized) ErrorAciertosEdadGenero=H-20s
 +       0.7477 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0.5177 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.4313 * (normalized) ErrorAciertosEdadGenero=Sin
```

```
 +       2.1506

Number of kernel evaluations: 3909278 (65.802% cached)
Classifier for classes: Male-10s, Female-30s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

        -0.6645 * (normalized) UsoPalabrasCerradas
 +      -0.0012 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.0006 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.0006 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.2228 * (normalized) RiquezaVocabulario=Pobre
 +      -0.222  * (normalized) RiquezaVocabulario=Promedio
 +       0.4448 * (normalized) RiquezaVocabulario=Rico
 +      -0.6676 * (normalized) SignosPuntuacion
 +      -0.221  * (normalized) LongitudMensaje=Corto
 +      -0.2221 * (normalized) LongitudMensaje=Medio
 +       0.4431 * (normalized) LongitudMensaje=Largo
 +       0.0002 * (normalized) ArgotsInternet=No-Usa
 +      -0.0002 * (normalized) ArgotsInternet=Baja
 +       0      * (normalized) ArgotsInternet=Alta
 +       0.6662 * (normalized) Emoticones=No-Usa
 +      -0.6644 * (normalized) Emoticones=Baja
 +      -0.0018 * (normalized) Emoticones=Alta
 +      -0.6656 * (normalized) TriEdad=Bajo
 +      -0.0005 * (normalized) TriEdad=Medio
 +       0.6646 * (normalized) TriEdad=Alto
 +       0.0015 * (normalized) TriEdad=Null
 +       0.1337 * (normalized) TriEdadGenero=TF10
 +       0.1338 * (normalized) TriEdadGenero=TF30
 +      -0.5331 * (normalized) TriEdadGenero=TM10
 +       0.1325 * (normalized) TriEdadGenero=TM20
 +       0      * (normalized) TriEdadGenero=TM30
 +       0.1332 * (normalized) TriEdadGenero=Null
 +       0.0004 * (normalized) TriGenero=fem
 +       0.0004 * (normalized) TriGenero=mal
 +      -0.0008 * (normalized) TriGenero=sin
 +       0.8887 * (normalized) ErrorGenero=F
 +      -1.1109 * (normalized) ErrorGenero=M
 +       0.2222 * (normalized) ErrorGenero=S
 +      -1.3321 * (normalized) ErrorEdad=E-10s
 +      -0.0002 * (normalized) ErrorEdad=E-20s
 +       1.3327 * (normalized) ErrorEdad=E-30s
 +      -0.0004 * (normalized) ErrorEdad=Sin
 +      -1.1658 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-20s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-30s
 +       0.8321 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0.8341 * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.5004 * (normalized) ErrorAciertosEdadGenero=Sin
 +       1.5883

Number of kernel evaluations: 1381436 (75.132% cached)

Classifier for classes: Male-20s, Male-30s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

         0.0002 * (normalized) UsoPalabrasCerradas
 +      -0.0002 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.0001 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.0001 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +       0      * (normalized) RiquezaVocabulario=Pobre
```

```
 +      -0.0001 * (normalized) RiquezaVocabulario=Promedio
 +       0.0001 * (normalized) RiquezaVocabulario=Rico
 +      -0.0002 * (normalized) SignosPuntuacion
 +       0.0003 * (normalized) LongitudMensaje=Corto
 +      -0.0002 * (normalized) LongitudMensaje=Medio
 +      -0.0001 * (normalized) LongitudMensaje=Largo
 +       0.0001 * (normalized) ArgotsInternet=No-Usa
 +       0      * (normalized) ArgotsInternet=Baja
 +      -0.0001 * (normalized) ArgotsInternet=Alta
 +       0.0003 * (normalized) Emoticones=No-Usa
 +      -0.0003 * (normalized) Emoticones=Baja
 +       0      * (normalized) Emoticones=Alta
 +       0      * (normalized) TriEdad=Bajo
 +      -0.0001 * (normalized) TriEdad=Medio
 +       0      * (normalized) TriEdad=Alto
 +       0.0001 * (normalized) TriEdad=Null
 +      -0.2855 * (normalized) TriEdadGenero=TF10
 +      -0.2859 * (normalized) TriEdadGenero=TF20
 +      -0.2855 * (normalized) TriEdadGenero=TF30
 +      -0.2856 * (normalized) TriEdadGenero=TM10
 +      -0.2856 * (normalized) TriEdadGenero=TM20
 +       1.7138 * (normalized) TriEdadGenero=TM30
 +      -0.2856 * (normalized) TriEdadGenero=Null
 +       0      * (normalized) TriGenero=fem
 +      -0.0001 * (normalized) TriGenero=mal
 +       0.0001 * (normalized) TriGenero=sin
 +       0      * (normalized) ErrorGenero=F
 +      -0.0001 * (normalized) ErrorGenero=M
 +       0.0001 * (normalized) ErrorGenero=S
 +      -0.9996 * (normalized) ErrorEdad=E-10s
 +      -0.9997 * (normalized) ErrorEdad=E-20s
 +       1.0002 * (normalized) ErrorEdad=E-30s
 +       0.9991 * (normalized) ErrorEdad=Sin
 +       0.2856 * (normalized) ErrorAciertosEdadGenero=H-10s
 +      -1.7128 * (normalized) ErrorAciertosEdadGenero=H-20s
 +       0.2859 * (normalized) ErrorAciertosEdadGenero=H-30s
 +       0.2855 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0.2851 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0.2853 * (normalized) ErrorAciertosEdadGenero=F-30s
 +       0.2854 * (normalized) ErrorAciertosEdadGenero=Sin
 -       0.0004

Number of kernel evaluations: 112431924 (30.023% cached)

Classifier for classes: Male-20s, Female-10s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

         0.5836 * (normalized) UsoPalabrasCerradas
 +       0.3016 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +      -0.0252 * (normalized) FrecuenciaPalabrasCerradas=Media
 +      -0.2764 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +       0.2221 * (normalized) RiquezaVocabulario=Pobre
 +       0.223  * (normalized) RiquezaVocabulario=Promedio
 +      -0.4451 * (normalized) RiquezaVocabulario=Rico
 +       0.2473 * (normalized) SignosPuntuacion
 +       0.2015 * (normalized) LongitudMensaje=Corto
 +       0.108  * (normalized) LongitudMensaje=Medio
 +      -0.3096 * (normalized) LongitudMensaje=Largo
 +      -0.0294 * (normalized) ArgotsInternet=No-Usa
 +      -0.0326 * (normalized) ArgotsInternet=Baja
 +       0.062  * (normalized) ArgotsInternet=Alta
 +      -0.4441 * (normalized) Emoticones=No-Usa
 +      -0.2741 * (normalized) Emoticones=Baja
```

```
 +       0.7182 * (normalized) Emoticones=Alta
 +       0.3899 * (normalized) TriEdad=Bajo
 +      -0.1896 * (normalized) TriEdad=Medio
 +      -0.1808 * (normalized) TriEdad=Alto
 +      -0.0195 * (normalized) TriEdad=Null
 +       1.1179 * (normalized) TriEdadGenero=TF10
 +       0      * (normalized) TriEdadGenero=TF20
 +       0.026  * (normalized) TriEdadGenero=TF30
 +       0.6178 * (normalized) TriEdadGenero=TM10
 +      -0.8814 * (normalized) TriEdadGenero=TM20
 +      -1      * (normalized) TriEdadGenero=TM30
 +       0.1197 * (normalized) TriEdadGenero=Null
 +       0.0024 * (normalized) TriGenero=fem
 +       0.0005 * (normalized) TriGenero=mal
 +      -0.0028 * (normalized) TriGenero=sin
 +       1.2245 * (normalized) ErrorGenero=F
 +      -1.1971 * (normalized) ErrorGenero=M
 +      -0.0273 * (normalized) ErrorGenero=S
 +       0.8577 * (normalized) ErrorEdad=E-10s
 +      -0.3933 * (normalized) ErrorEdad=E-20s
 +      -0.2319 * (normalized) ErrorEdad=E-30s
 +      -0.2324 * (normalized) ErrorEdad=Sin
 +      -0.5887 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-20s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-30s
 +       0.8327 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-20s
 +      -0.3274 * (normalized) ErrorAciertosEdadGenero=F-30s
 +       0.0835 * (normalized) ErrorAciertosEdadGenero=Sin
 -       2.0055

Number of kernel evaluations: 3927994 (67.288% cached)

Classifier for classes: Male-20s, Female-20s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

        -0.0006 * (normalized) UsoPalabrasCerradas
 +      -0.0001 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0      * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0      * (normalized) FrecuenciaPalabrasCerradas=Alta
 +       0      * (normalized) RiquezaVocabulario=Pobre
 +       0.0001 * (normalized) RiquezaVocabulario=Promedio
 +      -0.0001 * (normalized) RiquezaVocabulario=Rico
 +      -0.0008 * (normalized) SignosPuntuacion
 +      -0.0001 * (normalized) LongitudMensaje=Corto
 +       0.0001 * (normalized) LongitudMensaje=Medio
 +      -0.0001 * (normalized) LongitudMensaje=Largo
 +      -0.0001 * (normalized) ArgotsInternet=No-Usa
 +      -0.0002 * (normalized) ArgotsInternet=Baja
 +       0.0003 * (normalized) ArgotsInternet=Alta
 +       0.0002 * (normalized) Emoticones=No-Usa
 +      -0.0002 * (normalized) Emoticones=Baja
 +       0.0001 * (normalized) TriEdad=Bajo
 +       0      * (normalized) TriEdad=Medio
 +       0.0001 * (normalized) TriEdad=Alto
 +      -0.0002 * (normalized) TriEdad=Null
 +      -0.2853 * (normalized) TriEdadGenero=TF10
 +       1.713  * (normalized) TriEdadGenero=TF20
 +      -0.2853 * (normalized) TriEdadGenero=TF30
 +      -0.2854 * (normalized) TriEdadGenero=TM10
 +      -0.2862 * (normalized) TriEdadGenero=TM20
 +      -0.2851 * (normalized) TriEdadGenero=TM30
 +      -0.2857 * (normalized) TriEdadGenero=Null
```

```
 +       0.0003 * (normalized) TriGenero=fem
 +      -0.0002 * (normalized) TriGenero=mal
 +      -0.0001 * (normalized) TriGenero=sin
 +       1.3332 * (normalized) ErrorGenero=F
 +      -0.6676 * (normalized) ErrorGenero=M
 +      -0.6656 * (normalized) ErrorGenero=S
 +      -0.0002 * (normalized) ErrorEdad=E-10s
 +      -0.0001 * (normalized) ErrorEdad=E-20s
 +       0.0004 * (normalized) ErrorEdad=E-30s
 +       0      * (normalized) ErrorEdad=Sin
 +      -0.2849 * (normalized) ErrorAciertosEdadGenero=H-10s
 +      -0.2873 * (normalized) ErrorAciertosEdadGenero=H-20s
 +      -0.2849 * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -0.2853 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       1.7131 * (normalized) ErrorAciertosEdadGenero=F-20s
 +      -0.2855 * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.2851 * (normalized) ErrorAciertosEdadGenero=Sin
 +       0.2379

Number of kernel evaluations: 212074269 (26.609% cached)

Classifier for classes: Male-20s, Female-30s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

        -0.5819 * (normalized) UsoPalabrasCerradas
 +      -0.027  * (normalized) FrecuenciaPalabrasCerradas=Baja
 +      -0.027  * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.0539 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.0837 * (normalized) RiquezaVocabulario=Pobre
 +       0.0003 * (normalized) RiquezaVocabulario=Promedio
 +       0.0834 * (normalized) RiquezaVocabulario=Rico
 +      -0.664  * (normalized) SignosPuntuacion
 +       0.0828 * (normalized) LongitudMensaje=Corto
 +      -0.0007 * (normalized) LongitudMensaje=Medio
 +      -0.0822 * (normalized) LongitudMensaje=Largo
 +       0.3595 * (normalized) ArgotsInternet=No-Usa
 +       0.2769 * (normalized) ArgotsInternet=Baja
 +      -0.6364 * (normalized) ArgotsInternet=Alta
 +       0.2926 * (normalized) Emoticones=No-Usa
 +      -0.2926 * (normalized) Emoticones=Baja
 +      -0.1224 * (normalized) TriEdad=Bajo
 +      -0.3741 * (normalized) TriEdad=Medio
 +       0.2909 * (normalized) TriEdad=Alto
 +       0.2056 * (normalized) TriEdad=Null
 +       0.441  * (normalized) TriEdadGenero=TF10
 +       0.274  * (normalized) TriEdadGenero=TF20
 +       0.191  * (normalized) TriEdadGenero=TF30
 +       0.0221 * (normalized) TriEdadGenero=TM10
 +      -1.1424 * (normalized) TriEdadGenero=TM20
 +       0.1077 * (normalized) TriEdadGenero=TM30
 +       0.1068 * (normalized) TriEdadGenero=Null
 +      -0.0004 * (normalized) TriGenero=fem
 +      -0.0005 * (normalized) TriGenero=mal
 +       0.001  * (normalized) TriGenero=sin
 +       1.304  * (normalized) ErrorGenero=F
 +      -1.027  * (normalized) ErrorGenero=M
 +      -0.277  * (normalized) ErrorGenero=S
 +      -0.4585 * (normalized) ErrorEdad=E-10s
 +      -0.9582 * (normalized) ErrorEdad=E-20s
 +       1.1244 * (normalized) ErrorEdad=E-30s
 +       0.2923 * (normalized) ErrorEdad=Sin
 +       0.1087 * (normalized) ErrorAciertosEdadGenero=H-10s
 +      -0.1419 * (normalized) ErrorAciertosEdadGenero=H-20s
```

```
 +      -0.6406 * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -0.0607 * (normalized) ErrorAciertosEdadGenero=F-10s
 +      -0.0618 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0.7732 * (normalized) ErrorAciertosEdadGenero=F-30s
 +       0.0231 * (normalized) ErrorAciertosEdadGenero=Sin
 -       0.2285

Number of kernel evaluations: 153240673 (32.32% cached)

Classifier for classes: Male-30s, Female-10s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

         0      * (normalized) UsoPalabrasCerradas
 +       0.3823 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +      -0.1911 * (normalized) FrecuenciaPalabrasCerradas=Media
 +      -0.1912 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +       0.3794 * (normalized) RiquezaVocabulario=Pobre
 +       0.0962 * (normalized) RiquezaVocabulario=Promedio
 +      -0.4756 * (normalized) RiquezaVocabulario=Rico
 +      -0.0003 * (normalized) SignosPuntuacion
 +       0.0467 * (normalized) LongitudMensaje=Corto
 +       0.0476 * (normalized) LongitudMensaje=Medio
 +      -0.0943 * (normalized) LongitudMensaje=Largo
 +      -0.0711 * (normalized) ArgotsInternet=No-Usa
 +       0.0711 * (normalized) ArgotsInternet=Baja
 +       0      * (normalized) ArgotsInternet=Alta
 +      -0.3331 * (normalized) Emoticones=No-Usa
 +       0.0961 * (normalized) Emoticones=Baja
 +       0.237  * (normalized) Emoticones=Alta
 +       0.5364 * (normalized) TriEdad=Bajo
 +      -0.036  * (normalized) TriEdad=Medio
 +      -0.3217 * (normalized) TriEdad=Alto
 +      -0.1788 * (normalized) TriEdad=Null
 +       0.951  * (normalized) TriEdadGenero=TF10
 +      -0.762  * (normalized) TriEdadGenero=TF20
 +       0.0966 * (normalized) TriEdadGenero=TF30
 +       0.8099 * (normalized) TriEdadGenero=TM10
 +      -0.3337 * (normalized) TriEdadGenero=TM20
 +      -1      * (normalized) TriEdadGenero=TM30
 +       0.2382 * (normalized) TriEdadGenero=Null
 +       0.1904 * (normalized) TriGenero=fem
 +      -0.2381 * (normalized) TriGenero=mal
 +       0.0477 * (normalized) TriGenero=sin
 +       0.9523 * (normalized) ErrorGenero=F
 +      -0.9048 * (normalized) ErrorGenero=M
 +      -0.0475 * (normalized) ErrorGenero=S
 +       1.1791 * (normalized) ErrorEdad=E-10s
 +       0.1792 * (normalized) ErrorEdad=E-20s
 +      -1.1062 * (normalized) ErrorEdad=E-30s
 +      -0.2521 * (normalized) ErrorEdad=Sin
 +      -0.5396 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-20s
 +      -0.4018 * (normalized) ErrorAciertosEdadGenero=H-30s
 +       1.1718 * (normalized) ErrorAciertosEdadGenero=F-10s
 +      -0.5437 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-30s
 +       0.3134 * (normalized) ErrorAciertosEdadGenero=Sin
 -       2.2448

Number of kernel evaluations: 1249493 (74.721% cached)

Classifier for classes: Male-30s, Female-20s
BinarySMO
```

```
Machine linear: showing attribute weights, not support vectors.

        -0.558  * (normalized) UsoPalabrasCerradas
 +      -0.0502 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +      -0.0299 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.0801 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +       0.1485 * (normalized) RiquezaVocabulario=Pobre
 +       0.1487 * (normalized) RiquezaVocabulario=Promedio
 +      -0.2971 * (normalized) RiquezaVocabulario=Rico
 +      -0.1119 * (normalized) SignosPuntuacion
 +      -0.0662 * (normalized) LongitudMensaje=Corto
 +      -0.0001 * (normalized) LongitudMensaje=Medio
 +       0.0663 * (normalized) LongitudMensaje=Largo
 +      -0.1491 * (normalized) ArgotsInternet=No-Usa
 +      -0.0371 * (normalized) ArgotsInternet=Baja
 +       0.1862 * (normalized) ArgotsInternet=Alta
 +      -0.7339 * (normalized) Emoticones=No-Usa
 +       0.7339 * (normalized) Emoticones=Baja
 +      -0.0117 * (normalized) TriEdad=Bajo
 +       0.3007 * (normalized) TriEdad=Medio
 +      -0.1449 * (normalized) TriEdad=Alto
 +      -0.1441 * (normalized) TriEdad=Null
 +      -0.0236 * (normalized) TriEdadGenero=TF10
 +       2.2025 * (normalized) TriEdadGenero=TF20
 +      -0.023  * (normalized) TriEdadGenero=TF30
 +      -0.0229 * (normalized) TriEdadGenero=TM10
 +      -0.1104 * (normalized) TriEdadGenero=TM20
 +      -1.6007 * (normalized) TriEdadGenero=TM30
 +      -0.4221 * (normalized) TriEdadGenero=Null
 +       0.377  * (normalized) TriGenero=fem
 +      -0.2216 * (normalized) TriGenero=mal
 +      -0.1554 * (normalized) TriGenero=sin
 +       1.0229 * (normalized) ErrorGenero=F
 +      -1.0892 * (normalized) ErrorGenero=M
 +       0.0662 * (normalized) ErrorGenero=S
 +       0.3167 * (normalized) ErrorEdad=E-10s
 +       0.7162 * (normalized) ErrorEdad=E-20s
 +      -0.8381 * (normalized) ErrorEdad=E-30s
 +      -0.1948 * (normalized) ErrorEdad=Sin
 +       0.1825 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0.1401 * (normalized) ErrorAciertosEdadGenero=H-20s
 +      -1.0615 * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -0.104  * (normalized) ErrorAciertosEdadGenero=F-10s
 +       1.0753 * (normalized) ErrorAciertosEdadGenero=F-20s
 +      -0.1283 * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.1041 * (normalized) ErrorAciertosEdadGenero=Sin
 +       1.3632

Number of kernel evaluations: 109616904 (39.635% cached)

Classifier for classes: Male-30s, Female-30s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

        -1.0008 * (normalized) UsoPalabrasCerradas
 +       0.0002 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +      -0.0001 * (normalized) FrecuenciaPalabrasCerradas=Media
 +      -0.0002 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +       0.0003 * (normalized) RiquezaVocabulario=Pobre
 +      -0.0001 * (normalized) RiquezaVocabulario=Promedio
 +      -0.0002 * (normalized) RiquezaVocabulario=Rico
 +      -0.0007 * (normalized) SignosPuntuacion
 +      -0.0001 * (normalized) LongitudMensaje=Corto
 +      -0.0001 * (normalized) LongitudMensaje=Medio
```

```
 +       0.0002 * (normalized) LongitudMensaje=Largo
 +       0.0003 * (normalized) ArgotsInternet=No-Usa
 +       0      * (normalized) ArgotsInternet=Baja
 +      -0.0003 * (normalized) ArgotsInternet=Alta
 +      -0.5004 * (normalized) Emoticones=No-Usa
 +       0.5004 * (normalized) Emoticones=Baja
 +       0.0002 * (normalized) TriEdad=Bajo
 +      -0.0002 * (normalized) TriEdad=Medio
 +       0.0001 * (normalized) TriEdad=Alto
 +      -0.0001 * (normalized) TriEdad=Null
 +       0.8553 * (normalized) TriEdadGenero=TF10
 +      -0.1414 * (normalized) TriEdadGenero=TF20
 +       0.8566 * (normalized) TriEdadGenero=TF30
 +      -0.1426 * (normalized) TriEdadGenero=TM10
 +      -0.1426 * (normalized) TriEdadGenero=TM20
 +      -1.1428 * (normalized) TriEdadGenero=TM30
 +      -0.1425 * (normalized) TriEdadGenero=Null
 +       0.0003 * (normalized) TriGenero=fem
 +      -0.0002 * (normalized) TriGenero=mal
 +      -0.0001 * (normalized) TriGenero=sin
 +       0.9998 * (normalized) ErrorGenero=F
 +      -1.0003 * (normalized) ErrorGenero=M
 +       0.0005 * (normalized) ErrorGenero=S
 +       0.0001 * (normalized) ErrorEdad=E-10s
 +      -0.0002 * (normalized) ErrorEdad=E-20s
 +       0.0001 * (normalized) ErrorEdad=E-30s
 +       0      * (normalized) ErrorEdad=Sin
 +       0.0001 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0.0014 * (normalized) ErrorAciertosEdadGenero=H-20s
 +      -1.0012 * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -0.0002 * (normalized) ErrorAciertosEdadGenero=F-10s
 +      -0.0003 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       1.0002 * (normalized) ErrorAciertosEdadGenero=F-30s
 +       0      * (normalized) ErrorAciertosEdadGenero=Sin
 +       0.6431

Number of kernel evaluations: 102792260 (33.46% cached)

Classifier for classes: Female-10s, Female-20s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

         0      * (normalized) UsoPalabrasCerradas
 +      -0.2419 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.0303 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.2116 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.3333 * (normalized) RiquezaVocabulario=Pobre
 +      -0.0605 * (normalized) RiquezaVocabulario=Promedio
 +       0.3938 * (normalized) RiquezaVocabulario=Rico
 +      -0.0922 * (normalized) SignosPuntuacion
 +      -0.2123 * (normalized) LongitudMensaje=Corto
 +      -0.0303 * (normalized) LongitudMensaje=Medio
 +       0.2426 * (normalized) LongitudMensaje=Largo
 +      -0.0912 * (normalized) ArgotsInternet=No-Usa
 +      -0.0907 * (normalized) ArgotsInternet=Baja
 +       0.1819 * (normalized) ArgotsInternet=Alta
 +       0.5454 * (normalized) Emoticones=No-Usa
 +       0.2738 * (normalized) Emoticones=Baja
 +      -0.8192 * (normalized) Emoticones=Alta
 +      -0.424  * (normalized) TriEdad=Bajo
 +       0.3933 * (normalized) TriEdad=Medio
 +       0.0307 * (normalized) TriEdad=Alto
 +       0      * (normalized) TriEdad=Null
 +      -0.879  * (normalized) TriEdadGenero=TF10
```

```
 +       0.5765 * (normalized) TriEdadGenero=TF20
 +       0.3035 * (normalized) TriEdadGenero=TF30
 +      -0.6052 * (normalized) TriEdadGenero=TM10
 +       0.5735 * (normalized) TriEdadGenero=TM20
 +       0      * (normalized) TriEdadGenero=TM30
 +       0.0307 * (normalized) TriEdadGenero=Null
 +       0.1208 * (normalized) TriGenero=fem
 +      -0.0606 * (normalized) TriGenero=mal
 +      -0.0602 * (normalized) TriGenero=sin
 +      -0.2425 * (normalized) ErrorGenero=F
 +       0.2118 * (normalized) ErrorGenero=M
 +       0.0307 * (normalized) ErrorGenero=S
 +      -0.9777 * (normalized) ErrorEdad=E-10s
 +       0.7499 * (normalized) ErrorEdad=E-20s
 +       0.2946 * (normalized) ErrorEdad=E-30s
 +      -0.0669 * (normalized) ErrorEdad=Sin
 +       0.727  * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -1.0002 * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0.4561 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0      * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.1828 * (normalized) ErrorAciertosEdadGenero=Sin
 +       2.008

Number of kernel evaluations: 6780224 (63.237% cached)

Classifier for classes: Female-10s, Female-30s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

         0      * (normalized) UsoPalabrasCerradas
 +      -0.5461 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +       0.054  * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0.4922 * (normalized) FrecuenciaPalabrasCerradas=Alta
 +      -0.4611 * (normalized) RiquezaVocabulario=Pobre
 +      -0.3036 * (normalized) RiquezaVocabulario=Promedio
 +       0.7647 * (normalized) RiquezaVocabulario=Rico
 +      -0.646  * (normalized) SignosPuntuacion
 +      -0.2179 * (normalized) LongitudMensaje=Corto
 +      -0.1561 * (normalized) LongitudMensaje=Medio
 +       0.374  * (normalized) LongitudMensaje=Largo
 +       0.4075 * (normalized) ArgotsInternet=No-Usa
 +       0.1836 * (normalized) ArgotsInternet=Baja
 +      -0.5911 * (normalized) ArgotsInternet=Alta
 +       0.3006 * (normalized) Emoticones=No-Usa
 +      -0.0987 * (normalized) Emoticones=Baja
 +      -0.2019 * (normalized) Emoticones=Alta
 +      -0.6031 * (normalized) TriEdad=Bajo
 +       0.0476 * (normalized) TriEdad=Medio
 +       0.507  * (normalized) TriEdad=Alto
 +       0.0485 * (normalized) TriEdad=Null
 +      -1.0224 * (normalized) TriEdadGenero=TF10
 +       0.5405 * (normalized) TriEdadGenero=TF20
 +       0.2    * (normalized) TriEdadGenero=TF30
 +      -0.7382 * (normalized) TriEdadGenero=TM10
 +       0.3321 * (normalized) TriEdadGenero=TM20
 +       0.6965 * (normalized) TriEdadGenero=TM30
 +      -0.0085 * (normalized) TriEdadGenero=Null
 +       0.0408 * (normalized) TriGenero=fem
 +       0.1318 * (normalized) TriGenero=mal
 +      -0.1726 * (normalized) TriGenero=sin
 +      -0.34   * (normalized) ErrorGenero=F
 +       0.2166 * (normalized) ErrorGenero=M
 +       0.1234 * (normalized) ErrorGenero=S
```

```
 +      -1.2069 * (normalized) ErrorEdad=E-10s
 +       0.1157 * (normalized) ErrorEdad=E-20s
 +       1.0782 * (normalized) ErrorEdad=E-30s
 +       0.013  * (normalized) ErrorEdad=Sin
 +       0.9204 * (normalized) ErrorAciertosEdadGenero=H-10s
 +       0      * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -1.42   * (normalized) ErrorAciertosEdadGenero=F-10s
 +       0.4531 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       0.2971 * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.2507 * (normalized) ErrorAciertosEdadGenero=Sin
 +       2.4148

Number of kernel evaluations: 2381198 (71.353% cached)

Classifier for classes: Female-20s, Female-30s
BinarySMO
Machine linear: showing attribute weights, not support vectors.

        -0.0015 * (normalized) UsoPalabrasCerradas
 +       0.0001 * (normalized) FrecuenciaPalabrasCerradas=Baja
 +      -0.0001 * (normalized) FrecuenciaPalabrasCerradas=Media
 +       0      * (normalized) FrecuenciaPalabrasCerradas=Alta
 +       0      * (normalized) RiquezaVocabulario=Pobre
 +      -0.0001 * (normalized) RiquezaVocabulario=Promedio
 +       0.0001 * (normalized) RiquezaVocabulario=Rico
 +       0.0001 * (normalized) SignosPuntuacion
 +       0.0002 * (normalized) LongitudMensaje=Corto
 +      -0.0002 * (normalized) LongitudMensaje=Medio
 +       0      * (normalized) LongitudMensaje=Largo
 +       0.0001 * (normalized) ArgotsInternet=No-Usa
 +       0      * (normalized) ArgotsInternet=Baja
 +      -0.0001 * (normalized) ArgotsInternet=Alta
 +       0.0001 * (normalized) Emoticones=No-Usa
 +      -0.0001 * (normalized) Emoticones=Baja
 +      -0.0001 * (normalized) Emoticones=Alta
 +      -0.0003 * (normalized) TriEdad=Bajo
 +      -0.0001 * (normalized) TriEdad=Medio
 +       0.0002 * (normalized) TriEdad=Alto
 +       0.0002 * (normalized) TriEdad=Null
 +       0.2857 * (normalized) TriEdadGenero=TF10
 +      -1.7147 * (normalized) TriEdadGenero=TF20
 +       0.2858 * (normalized) TriEdadGenero=TF30
 +       0.2858 * (normalized) TriEdadGenero=TM10
 +       0.2856 * (normalized) TriEdadGenero=TM20
 +       0.2858 * (normalized) TriEdadGenero=TM30
 +       0.286  * (normalized) TriEdadGenero=Null
 +       0.0001 * (normalized) TriGenero=fem
 +      -0.0001 * (normalized) TriGenero=mal
 +       0      * (normalized) TriGenero=sin
 +       0.0001 * (normalized) ErrorGenero=F
 +      -0.0002 * (normalized) ErrorGenero=M
 +       0.0001 * (normalized) ErrorGenero=S
 +      -0.9996 * (normalized) ErrorEdad=E-10s
 +      -0.9998 * (normalized) ErrorEdad=E-20s
 +       1.0002 * (normalized) ErrorEdad=E-30s
 +       0.9992 * (normalized) ErrorEdad=Sin
 +      -0.2857 * (normalized) ErrorAciertosEdadGenero=H-10s
 +      -0.2849 * (normalized) ErrorAciertosEdadGenero=H-20s
 +      -0.2857 * (normalized) ErrorAciertosEdadGenero=H-30s
 +      -0.2856 * (normalized) ErrorAciertosEdadGenero=F-10s
 +      -0.2859 * (normalized) ErrorAciertosEdadGenero=F-20s
 +       1.7135 * (normalized) ErrorAciertosEdadGenero=F-30s
 +      -0.2856 * (normalized) ErrorAciertosEdadGenero=Sin
 -       0.0007
```

```
Number of kernel evaluations: 131167658 (29.285% cached)

Time taken to build model: 349.58 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13782               68.91   %
Incorrectly Classified Instances      6218               31.09   %
Kappa statistic                          0.5879
Mean absolute error                      0.2345
Root mean squared error                  0.3294
Relative absolute error                 92.6777 %
Root relative squared error             92.5965 %
Total Number of Instances            20000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.554     0.006      0.617     0.554     0.584      0.904    Male-10s
                 0.814     0.195      0.625     0.814     0.707      0.828    Male-20s
                 0.619     0.049      0.757     0.619     0.681      0.839    Male-30s
                 0.431     0.004      0.67      0.431     0.524      0.91     Female-10s
                 0.661     0.063      0.806     0.661     0.726      0.835    Female-20s
                 0.654     0.097      0.628     0.654     0.641      0.813    Female-30s
Weighted Avg.    0.689     0.103      0.703     0.689     0.689      0.832

=== Confusion Matrix ===

    a    b    c    d    e    f   <-- classified as
  184  114    9    2    5   18 |    a = Male-10s
   49 4644  305   10  316  381 |    b = Male-20s
   34  784 2454    1   95  595 |    c = Male-30s
    2   76    5  152  106   12 |    d = Female-10s
   17 1175  129   47 3728  547 |    e = Female-20s
   12  640  341   15  376 2620 |    f = Female-30s
```

Con estos resultados, se procede a hacer una tabla comparativa entre los tres algoritmos para identificar al clasificador que nos aporta un mayor porcentaje de efectividad en el menor tiempo.

Tabla 2. Comparativa de efectividad por Clasificador (para Edad).

CLASIFICADOR	PREDICCIÓN DE EDAD	TIEMPO DE CONSTRUCCIÓN DEL MODELO
J48 (C4.5)	69.3 %	0.13 seg.
Naive Bayes	69.34 %	0.07 seg.
Maquinas de Soporte Vectorial	68.91 %	349.58 seg.

CAPITULO 4. CONCLUSIONES

A lo largo de este trabajo hemos experimentado con tres distintos algoritmos de clasificación, con el propósito de predecir la edad y género en textos en español, a través de la identificación de características determinantes, que nos ayuden a una correcta clasificación.

Se a visto que la elección correcta de características discriminantes, son el paso más importante en esta investigación. Se observó que, a medida que se identifican e implementan mejores características, se obtuvieron mejores resultados. Por otra parte, la implementación de los tres algoritmos ayudo a poder obtener información base para este trabajo, que posteriormente nos ayudaría a medir el porcentaje de efectividad en el que nos encontramos en cada una de las etapas de este proyecto.

Con la siguiente tabla se muestran las comparativas y el crecimiento del porcentaje de efectividad en cuanto a la correcta clasificación, de acuerdo como se fueron integrando las nuevas características al proyecto. Mostrando así que las características individualmente aportan un porcentaje bajo mientras que trabajando en conjunto fueron determinantes para ayudar a la clasificación y en particular las características “Trigramas” y “Errores y Aciertos” fueron quienes ayudaron con un porcentaje mayor.

Tabla 3. Comparativa de efectividad por Clasificador (para Edad - Género).

<table>
<tr><th colspan="3">% Conjunto</th><th>Características</th><th>Porcentaje Individual</th></tr>
<tr><td rowspan="9">81.32%</td><td rowspan="8">72.012%</td><td rowspan="7">58.58%</td><td>Uso de Palabras Cerradas</td><td>50.29 %</td></tr>
<tr><td>Frecuencia de Palabras Cerradas</td><td>52.98 %</td></tr>
<tr><td>Riqueza del Vocabulario</td><td>51.79 %</td></tr>
<tr><td>Signos de puntuación</td><td>51.89 %</td></tr>
<tr><td>Longitud del Mensaje</td><td>53.78 %</td></tr>
<tr><td>Argots de Internet</td><td>52.19 %</td></tr>
<tr><td>Emoticones</td><td>50. 29 %</td></tr>
<tr><td colspan="2">Trigramas</td><td>64.41 %</td></tr>
<tr><td colspan="3">Errores y Edades</td><td>67.71 %</td></tr>
</table>

Para concluir se muestra en la siguiente tabla la comparativa en cuanto a porcentaje de efectividad y tiempo de construcción del modelo de clasificación, con lo que se puede decidir que clasificador utilizar, para cada caso que se requiera predecir. Como se puede observar el algoritmo J48 fue el que mejor se comportó a la hora de clasificar por separado y cuando se clasifico en conjunto el algoritmo Naive Bayes supero con un porcentaje mínimo pero con un menor tiempo al algoritmo J48 que venía comportándose mejor en las pruebas de clasificación individuales.

Tabla 4. Comparativa de efectividad por Clasificador (para Género, Edad y Edad - Género).

	CLASIFICADOR	**PORCENTAJE DE EFECTIVIDAD**	**TIEMPO DE CONSTRUCCION DEL MODELO**
GÉNERO	J48 (C4.5)	**82.71 %**	0.11 seg.
	Naive Bayes	82.325%	**0.03 seg.**
	Maquinas de Soporte Vectorial	82.045%	341.03 seg.
EDAD	J48 (C4.5)	**81.09 %**	0.34 seg.
	Naive Bayes	80.67 %	**0.03 seg.**
	Maquinas de Soporte Vectorial	80.38 %	257.28 seg.
EDAD y GÉNERO	J48 (C4.5)	69.3 %	0.13 seg.
	Naive Bayes	**69.34 %**	**0.07 seg.**
	Maquinas de Soporte Vectorial	68.07 %	349.58 seg.

REFERENCIAS BIBLIOGRAFICAS

[1] Ozcan ozyurt,cemal kose.,2010,"Chat mining:Automatically Determination of Chat Conversation‟s Topic in text based chat mediums",Journal on selected areas in Expert Systems with Applications (2010) ESWA 4843,www.elsevier.com/locate/eswa do:10,1016/j.eswa,2010.06,053.

[2] C.Rose, O.Ozyurt and G.Amanmyradov..,2007,"Mining Chat Conversations for Sex Identification",Proceedings of the IEEE International Conference on Statistical and Semantic approaches for sex identification,PAKDD 2007 Workshops,LNAI 4819,PP.45-55.

[3] Haichao, D., Siu, C.H., Yulan, H.: Structural analysis of chat messages for topic detection.Online Information Review 30(5), 496–516 (2006).

[4] Vel, O. de, Corney, M., Anderson, A., Mohay, G.: Language and Gender Author Cohort Analysis of E-mail for Computer Forensics. In: Second Digital Forensics Research Workshop. (2002).

[5] Han, E., Karypis, G., & Kumar, V. (2001). Text categorization using weight adjusted k-nearest neighbor classification. Lecture Notes in Computer Science, 2035, 53–65.

[6] Shanmugasundaram Hariharan and Aashika Rani.K. March 2011., "Gender Prediction in Chat based Medium's Using Text Mining" International Journal of Research and Reviews in Information Sciences

[7] Corney, M.W.: Analysing E-mail Text Authorship for Forensic Purposes. M.S. Thesis. Queensland University of Technology (2003)

[8] Holmes, I., Forstyh, R.: The Federalist Revisited: New Directions in Authorship Attribution. Literary and Linguistic Computing 10(2) (1995) 111–127

[9] Graham, N., Hirst, G., Marthi, B.: Segmenting Documents by Stylistic Character. Natural Language Engineering 11(4) (2005) 397–415

[10] Koppel, M., Argamon, S., Shimoni, A.R.: Automatically Categorizing Written Texts by Author Gender. Literary & Linguistic Computing 17(4) (2002) 401–412

[11] T. Kucukyilmaz, B. Barla Cambazoglu, C. Aykanat, and F. Can.: Chat Mining for Gender Prediction. Bilkent University, Department of Computer Engineering (2006)

[12] http://pan.webis.de Fecha de consulta Mayo 2013

[13] http://www.clef-initiative.eu// Fecha de consulta Mayo 2013

[14] http://es.scribd.com/doc/21458936/80/Clasificadores-supervisados-y-no-supervisados Capítulo 5 –Introducción a los clasificadores - 178 – Mayo 2013 - José Francisco Vélez Serrano, Ana Belén Moreno Díaz, Ángel Sánchez Calle, José Luis Esteban Sánchez-Marín

[15] ING. BRUNO LÓPEZ TAKEYAS Inteligencia Artificial Descripción del "Algoritmo C4.5" Instituto Tecnológico de Nuevo Laredo, Tamaulipas, Noviembre del 2005.

[16] Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid "Tema 6. Clasificadores Bayesianos". Departamento de Ciencias de la Computación e Inteligencia Artificial – Universidad del País Vasco-Euskal Herriko Unibertsitatea (1997).

[17] GUSTAVO A. BETANCOURT "LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs)" Scientia et Technica Año XI, No 27, Abril 2005. UTP. ISSN 0122-1701