METODOLÓGICO

Documentos de trabajo

Clasificación automática de textos utilizando técnicas de *text mining*: Aplicación a las glosas de la Encuesta Nacional de Empleo (ENE)

Autores:
Julio Guerrero
Julián Cabezas

N°6, marzo de 2019

Instituto Nacional de Estadísticas · Chile

INSTITUTO NACIONAL DE ESTADÍSTICAS

Morandé 801, Santiago de Chile

Teléfono: 56 232461000

Correo: ine@ine.cl

Facebook: @ChileINE

Twitter: @INE_Chile

Julio Guerrero

Julián Cabezas

Subdirección Técnica

Los autores agradecen la colaboración del encargado de esta iniciativa, Jose Luis Aránguiz, de los analistas del Departamento de Estadísticas de Hogares, Nicolás Von Hausen y Nicolás Maturana, y de la Jefa del Departamento de Investigación y Desarrollo, Denisse López, cuyos comentarios e interés en la adopción de nuevas tecnologías hicieron posible este trabajo. También reconocen la participación y apoyo de compañeros pertenecientes al Departamento de Infraestructura Estadística, al Departamento de Estudios Laborales, al Departamento de TI, al Departamento de Presupuestos Familiares y al Departamento de Investigación y Desarrollo, quienes brindaron muchas horas de trabajo e insumos base que contribuyeron en el éxito de este trabajo.

Los Documentos de Trabajo del INE están dirigidos a investigadores, académicos, estudiantes y público especializado en materias económicas, y tienen como objetivo proporcionar un análisis exhaustivo sobre aspectos conceptuales, analíticos y metodológicos claves de los productos estadísticos que elabora la institución y, de esta forma, contribuir al intercambio de ideas entre los distintos componentes del Sistema Estadístico Nacional.

Las interpretaciones y opiniones que se expresan en los Documentos de Trabajo pertenecen en forma exclusiva a los autores y colaboradores y no reflejan necesariamente el punto de vista oficial del INE ni de la institución a la que pertenecen los colaboradores de los documentos.

El uso de un lenguaje que no discrimine ni marque diferencias entre hombres y mujeres ha sido una preocupación en la elaboración de este documento. Sin embargo, y con el fin de evitar la sobrecarga gráfica que supondría utilizar en castellano “o/a” para marcar la existencia de ambos sexos, se ha optado por utilizar -en la mayor parte de los casos- el masculino genérico, en el entendido de que todas las menciones en tal género representan siempre a hombres y mujeres, abarcando claramente ambos sexos.

Clasificación automática de textos utilizando técnicas de *text mining*: Aplicación a las glosas de la Encuesta Nacional de Empleo (ENE)

Resumen

Este documento explora los aspectos metodológicos relacionados con la clasificación automática de textos, tarea que consiste en asignar documentos de texto libres a una o más clases predefinidas, basadas en su contenido. Para este fin se describe la utilización de tres técnicas de *machine learning*: Naïve Bayes (NB), Support Vector Machine (SVM) y Random Forests (RF). Este estudio analiza las propiedades particulares del aprendizaje con datos de texto e identifica por qué estas técnicas son apropiadas para esta tarea. Una evaluación empírica de estas técnicas se realizó para apoyar los hallazgos teóricos, considerando la clasificación del "oficio, labor u ocupación" y del "sector económico" de la población ocupada a partir de los datos de la Encuesta Nacional de Empleo (ENE), recopilados durante 2017 por el Instituto Nacional de Estadísticas (INE). Las tres técnicas evaluadas mostraron buen desempeño en la tarea de clasificación, siendo SVM la de mejor desempeño, con una precisión global de alrededor del 90%. SVM logra un comportamiento sólido en una variedad de diferentes aprendizajes y es completamente automático, eliminando la necesidad de ajuste manual de parámetros.

Abstract

This document explores the methodological aspects of the automatic classification of texts, a task that consists in the assignment of free text documents to one or more predefined classes based on their content. To this end, the use of three machine learning techniques are described: Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forests (RF). This study analyses the particular properties of learning with text data and identifies why these techniques are appropriate for this task. An empirical evaluation of these techniques was made to support the theoretical findings, considering the classification of the "profession, job or occupation" and the "economic sector" of the employed population based on data from the National Employment Survey (ENE), collected during 2017 by the National Statistics Institute (INE). All the evaluated techniques performed well in the classification; SVM, with a global precision of around 90%, peformed best. SVM shows consistent precision in a wide variety of situations and is completely automatic, eliminating the need for manual adjustment of parameters.

Palabras clave: *text mining*, clasificación de textos, *machine learning*, encuesta nacional de empleo

1. Introducción

Con el rápido crecimiento de la información en línea, la clasificación de textos se ha convertido en una de las técnicas clave para manejar y organizar los datos de texto. Las técnicas de categorización de texto se utilizan para clasificar las noticias, para encontrar información de interés en la World Wide Web (WWW) y para guiar la búsqueda de un usuario a través del hipertexto (Joachims, 1998). En este contexto se emplean profesionales entrenados para clasificar nuevos ítems. Este proceso es muy costoso y consume mucho tiempo, limitando su aplicabilidad. Consecuentemente, existe un interés creciente por el desarrollo de tecnologías para la clasificación automática de textos (Aas & Eikvil, 1999).

Para estos fines, una serie de técnicas de clasificación estadística y *machine learning* han sido aplicadas para la clasificación de textos, incluyendo modelos de regresión (Aas & Eikvil, 1999), clasificadores basados en vecinos más cercanos (Aas & Eikvil, 1999; Pérez, 2017), árboles de decisión (Pérez, 2017), clasificadores Bayesianos (Aas & Eikvil, 1999; Mitchell, 1997), algoritmos de aprendizaje de reglas (Aas & Eikvil, 1999), Support Vector Machine (Joachims, 1998; Berry & Kogan, 2010) y redes neuronales (Aas & Eikvil), entre otras.

Esta misma realidad se ha vivido en el Instituto Nacional de Estadísticas (INE). Este organismo, en el ejercicio de su rol público, tiene la necesidad de desarrollar e implementar herramientas tecnológicas que aborden las tareas de codificación de grandes volúmenes de textos provenientes de las preguntas abiertas en las encuestas que desarrolla, en el marco del proceso de producción requerido para la elaboración de estadísticas laborales y sociodemográficas, con plazos perentorios de publicación. Actualmente, el proceso de codificación es realizado de manera manual, con una precisión de aproximadamente 84%, e implica más de 3.600 horas efectivas de trabajo al mes. Por esta razón, es necesario el desarrollo de una solución que supere el desempeño de la clasificación manual en términos de eficiencia y velocidad.

Para evaluar qué solución se podría implementar en el INE, se decidió medir la calidad y el avance en el campo de la clasificación de textos, lo que requiere contar

con una colección estandarizada de documentos para análisis y pruebas (Aas & Eikvil, 1999). En ese marco, en este documento de trabajo se aplicó el procedimiento de clasificación automática de textos sobre un conjunto de datos que contiene las glosas (documentos) que pesquisan "oficio, labor u ocupación" y el "sector económico" de la población ocupada a partir de los datos de la Encuesta Nacional de Empleo (ENE), recopilados durante 2017 por el INE. Así, este estudio tiene como objetivos 1) describir los pasos inherentes a un proceso de clasificación automática de textos siguiendo el esquema de *text mining*; 2) analizar algoritmos apropiados para la clasificación de textos, esto es, algoritmos de *machine learning*, como el Naïve Bayes (NB), Support Vector Machine (SVM) y Random Forests (RF); 3) construir clasificadores automáticos para las glosas provenientes de la ENE basados en las técnicas anteriores, y 4) evaluar el desempeño de los clasificadores desarrollados mediante métricas estadísticas y seleccionar el más apropiado.

En este trabajo se analizarán los aspectos metodológicos relacionados con un problema de clasificación automática de textos, lo que comprende desde la transformación de textos hacia una representación adecuada para las tareas de clasificación -lo que ha sido tradicionalmente abordado usando un modelo de espacio vectorial debido a su simplicidad y buen desempeño (Aas & Eikvil, 1999; Alfaro & Allende, 2011; Welbers et al., 2017)-, hasta la aplicación de una técnica de clasificación y su posterior evaluación.

El documento está organizado en seis secciones: la primera, como se vio, corresponde a la presentación del estudio. En la número 2 se desarrollan los pasos necesarios para transformar textos en bruto en una representación adecuada para las tareas de clasificación. En la sección 3 se describen tres técnicas que han sido exitosamente utilizadas en la clasificación de textos y que han sido consideradas para los objetivos de este trabajo. En la sección 4 se introducen métricas de desempeño para la evaluación de la clasificación en un problema binario. En la 5 se muestran los experimentos realizados sobre el conjunto de datos de la ENE 2017. Finalmente, en la sección 6 se entregan conclusiones y proyecciones sobre los resultados obtenidos.

2. Preparación de datos

La preparación de la data es el punto de partida de cualquier análisis estadístico de datos. El análisis computacional de textos no solo no es diferente en este aspecto, sino que, además, presenta con frecuencia algunos desafíos especiales que pueden resultar desalentadores tanto para analistas principiantes como avanzados. A su vez, la preparación de textos para el análisis requiere tomar decisiones que pueden afectar la precisión, la validez y los hallazgos de un estudio de análisis de textos, así como las técnicas o métodos utilizados en el análisis (Welbers et al., 2017).

En esta sección se describen algunos procedimientos de preprocesamiento, así como criterios para la selección de características y representación de textos, para luego abordar las medidas de efectividad de clasificación para el filtrado de clases.

2.1. Preprocesamiento

El paso inicial en la clasificación de textos es transformar documentos que típicamente son cadenas de caracteres en una representación adecuada para el algoritmo de aprendizaje y para la tarea de clasificación. La transformación de textos procede de la siguiente forma (Aas & Eikvil, 1999; Welbers et al., 2017):

Primero se aplica el proceso de **normalización de textos,** es decir, se transforman palabras en un formato más uniforme, de manera que un clasificador pueda reconocer cuando dos palabras tienen (aproximadamente) el mismo significado, incluso si están escritas ligeramente diferentes. Esto permite reducir el tamaño del vocabulario (i.e. el rango completo o dimensión de las características utilizadas en el análisis). El proceso comprende las siguientes etapas, cuyo orden de ejecución no es arbitrario: (1) convertir todo el texto a minúscula, (2) eliminar los *html* u otros *tags* (caracteres especiales), (3) quitar los signos de puntuación, (4) sacar los números y (5) eliminar los espacios en blanco múltiples.

Luego tiene lugar la **tokenización**, proceso que consiste en dividir un texto o documento en características más específicas conocidas como *tokens*, que típicamente son palabras o combinaciones de palabras que constituyen los componentes semánticos más significativos de los textos. La tokenización es crucial para el análisis computacional, porque los textos completos son demasiado específicos para realizar cualquier cálculo significativo.

2.2. Selección de características

En la clasificación de textos, la selección de características apropiadas (una palabra o un *token* en un documento) puede ser bastante útil. Estas son seleccionadas de acuerdo con sus contribuciones para la discriminación entre clases; si no son seleccionadas, son eliminadas de los datos para el aprendizaje e implementación de modelos. Los objetivos de la selección de características son dos: reducir la dimensionalidad en el espacio de características del documento y filtrar las características irrelevantes, ayudando a construir un modelo preciso y eficaz para la clasificación de documentos.

La reducción de la dimensionalidad busca disminuir el número de características que se van a modelar mientras se conserva el contenido de los documentos individuales, lo que generalmente ayuda a acelerar el proceso de entrenamiento de un modelo. La filtración, en tanto, es valiosa para ciertos algoritmos de *machine learning*, como las redes neuronales RBF, que tratan cada característica de datos por igual en sus cálculos de distancia y, por lo tanto, son incapaces de distinguir entre características relevantes e irrelevantes.

2.2.1. Procedimiento

La selección de características contempla dos pasos (Berry & Kogan, 2010):

1. Para un conjunto determinado de datos, las características se extraen y seleccionan bajo un esquema no supervisado[1]. Esto se lleva a cabo eliminando palabras comunes o de alta frecuencia denominadas *stopwords*, que son palabras que no contienen información acerca del contenido del documento (i.e. pronombres, artículos, preposiciones, conjunciones, etc.), y aplicando un procedimiento de *stemming*, es decir, remover sufijos para generar "palabras tallos o de origen". Esto se realiza para agrupar palabras que tienen el mismo significado conceptual, como, por ejemplo, hacer y haciendo.

2. Luego, las características con baja frecuencia en los documentos o baja frecuencia en el *corpus* (i.e. colección de documentos) respecto un umbral definido, se eliminan del conjunto de datos, ya que estas características pueden no ser de ayuda en la diferenciación de los documentos por clases y pueden agregar ruido en la

[1] Un esquema no supervisado se refiere al aprendizaje automático, en el que un modelo es ajustado a las observaciones sin conocimiento *a priori* de las etiquetas o clases a las que pertenecen.

clasificación de estos. El proceso de selección también elimina aquellas características con frecuencias muy altas en el *corpus* en el conjunto de datos, ya que muchas de ellas se distribuyen casi por igual entre las distintas clases y puede que no sean valiosas para caracterizar las clases de las características. A continuación, las características se seleccionan por sus distribuciones de frecuencias entre los documentos de entrenamiento de las distintas clases. Este procedimiento de selección de características supervisado pretende, utilizando los documentos de entrenamiento etiquetados, identificar de mejor forma las características que tienen mayor poder de discriminación entre las clases.

2.2.2. Métodos para la selección de características

Existen varios métodos supervisados[2] para la selección de características que han sido ampliamente utilizados en clasificación de textos (Sebastiani, 2002). Entre ellos destacan las siguientes métricas: la ganancia de información (IG), la estadística Chi-Cuadrado (CHI) y la métrica Odds Ratio (OR).

Ganancia de Información (IG)

El criterio IG cuantifica la cantidad de información obtenida para la predicción de clases mediante el conocimiento de la presencia o ausencia de una característica en el documento. El IG de una característica t sobre una clase c puede expresarse como:

$$IG(t,c) = \sum_{c\in\{c,\bar{c}\}} \sum_{t\in\{t,\bar{t}\}} P(t,c)\, log\left(\frac{P(t,c)}{P(t)P(c)}\right) \qquad \textbf{(1)}$$

donde $P(c)$ y $P(t)$ denotan la probabilidad de que un documento pertenezca a la clase c y la probabilidad que una característica t ocurra en un documento, respectivamente. $P(t,c)$ denota la probabilidad conjunta de t y c.

[2] Un método supervisado es una técnica cuyo aprendizaje es realizado a partir de un conjunto de ejemplos (documentos) etiquetados sobre los que se tiene un conocimiento *a priori* respecto a la clase a la cual pertenecen.

Todas las probabilidades pueden estimarse mediante conteos de frecuencias a partir de los datos de entrenamiento.

Estadístico Chi-Cuadrado (CHI)

Otro popular método de selección de características es el estadístico CHI. Este estadístico mide la falta de independencia entre la ocurrencia de la característica t y la ocurrencia de la clase c. Las características se clasifican respecto a la siguiente cantidad:

$$CHI(t,c) = \frac{n[P(t,c)P(\bar{t},\bar{c}) - P(t,\bar{c})P(\bar{t},c)]^2}{P(t)P(\bar{t})P(c)P(\bar{c})} \quad \textbf{(2)}$$

donde n es el tamaño de la data de entrenamiento y las notaciones de probabilidad tienen la misma interpretación como en la ecuación (**1**). Por ejemplo, $P(\bar{c})$ representa la probabilidad de que un documento no pertenezca a la clase c.

Odds Ratio (OR)

El tercer criterio de selección de características, OR, también ha sido utilizado en la clasificación de textos y mide la razón entre la probabilidad de ocurrencia de la característica t en un documento de clase c y la probabilidad de que la característica no ocurre en c, y se puede definir como:

$$OR(t,c) = \frac{P(t \mid c)\left(1 - P(t \mid \bar{c})\right)}{\left(1 - P(t \mid c)\right)P(t \mid \bar{c})} \quad \textbf{(3)}$$

La efectividad de los métodos de selección de características para la clasificación de textos ha sido estudiada y comparada, por ejemplo, por Yang & Pedersen (1997).

Estos autores llegaron a la conclusión de que la ganancia de información (IG) produce los resultados más estables.

2.3. Representación de documentos

La representación de documento quizás más comúnmente utilizada por su simplicidad y efectividad es el llamado Modelo de Espacio Vectorial (Aas & Eikvil, 1999). En este, los documentos son representados por vectores de palabras. Usualmente, se tiene una colección de documentos que son representados por una matriz de documento – término denotada por $\boldsymbol{A}$, donde cada entrada representa las ocurrencias de una palabra en el documento, es decir,

$$\boldsymbol{A} = (a_{dt}) \tag{4}$$

donde a_{dt} es el peso de la palabra t en el documento d. Como cada palabra normalmente no aparece en cada documento, la matriz $\boldsymbol{A}$ es usualmente *una matriz dispersa* (i.e. es una matriz de gran tamaño en la que la mayor parte de sus elementos son cero). El número de columnas (N) de la matriz corresponde al número de palabras en el diccionario, por lo que puede ser un número muy grande.

La matriz $\boldsymbol{A}$ es uno de los formatos más comunes para representar el *corpus* de un texto en un formato de bolsa de palabras (BOW, por sus siglas en inglés). La ventaja de esta representación es que permite analizar los datos con álgebra de vectores y matrices, moviéndose efectivamente del texto a los números. Además, con el uso de formatos especiales de matrices para *matrices dispersas*, los datos de textos en un formato $\boldsymbol{A}$ son muy eficientes en memoria y pueden ser analizados con operaciones altamente optimizadas (Welbers et al., 2017).

Existen varias formas de determinar los pesos a_{dt} de la palabra t en el documento d (Aas & Eikvil, 1999), pero la mayoría de los enfoques se basan en dos observaciones empíricas respecto al texto:

1. Cuantas más veces una palabra aparece en un documento, más relevante es para el tema del documento.
2. Cuantas más veces una palabra aparece a lo largo de todos los documentos en la colección, más pobremente discrimina entre documentos.

Sea f_{dt} la frecuencia de la palabra t en el documento d, M es el número de documentos en la colección, N es el número de palabras en la colección después de que los *stopwords* han sido removidos y se ha realizado un procedimiento de *stemming* (véase sección 2.2.), y n_t el número total de veces en que la palabra t ocurre en toda la colección.

El enfoque más simple consiste en asignar un peso 1 si la palabra aparece en el documento y 0 en otro caso, es decir,

$$a_{dt} = \begin{cases} 1, & si\ f_{dt} > 0 \\ 0, & e.o.c. \end{cases} \quad (5)$$

Otro enfoque simple consiste en usar la frecuencia de la palabra en el documento,

$$a_{dt} = f_{dt} \quad (6)$$

Sin embargo, los dos esquemas anteriores no toman en cuenta la frecuencia de la palabra a través de todos los documentos en la colección. Un enfoque bien conocido para el cálculo de los pesos de las palabras es el peso de término '$tf - idf$', que asigna los pesos a la palabra t en el documento d en proporción al número de ocurrencias de la palabra en el documento, y en proporción inversa al número de documentos en la colección en los que la palabra ocurre al menos una vez. Este esquema de

ponderación de términos produce buenos resultados de clasificación y ha sido seleccionado para este trabajo. Su cálculo viene dado por la siguiente expresión:

$$a_{dt} = f_{dt} \cdot log\left(\frac{M}{n_t}\right) \quad (7)$$

3. Métodos de clasificación de textos

La clasificación de textos es el problema que consiste en asignar automáticamente una o más clases predefinidas a documentos de texto libre (Aas & Eikvil, 1999). En la medida en que se dispone de mayor volumen de información de textos, la efectiva recuperación es difícil sin una buena indexación y resumen del contenido de un documento. La clasificación de documentos es una solución a este problema. Un creciente número de métodos estadísticos y técnicas de *machine learning* para clasificación han sido aplicados a la clasificación de textos en los años recientes.

La mayor parte de la investigación en esta materia se ha dedicado a problemas binarios, en los que un documento es clasificado como *relevante* o *no relevante* con respecto a un tópico de interés predefinido. Sin embargo, hay muchas fuentes de datos textuales como noticias de internet, correo electrónico y bibliotecas digitales, que están compuestas de diferentes temas y que, por lo tanto, plantean un problema de clasificación de clases múltiples.

Un enfoque común para el problema de clasificación con clases múltiples es llevar la tarea hacia problemas disjuntos de clasificación binaria, una para cada clase. Particularmente, en el caso de métodos como Naïve Bayes y Support Vector Machine, para realizar la clasificación de un nuevo documento se requiere aplicar todos los clasificadores binarios y combinar sus predicciones hacia una única decisión. El resultado final es un ranking de posibles temas, elaborado según la probabilidad de pertenencia a cada clase.

En lo que sigue se describen algunos de los algoritmos para clasificación de textos que han sido propuestos y evaluados en el pasado y que han sido considerados como alternativas de clasificador en este trabajo, a saber: Naïve Bayes, Support Vector Machine y Random Forests. Cabe señalar que estos algoritmos operan bajo un esquema supervisado, es decir, el entrenamiento del clasificador se realiza utilizando ejemplos (documentos) cuya clase está previamente etiquetada.

En primer lugar se entrega alguna notación general: Sea $d = \{d_1, \ldots, d_m\}$ un vector de documentos para ser clasificado y $c_1, \ldots, c_k$ las posibles clases. Se asume que cada documento d_i puede ser expresado como un vector numérico representando los pesos de términos o características $d_i = \{t_1, \ldots, t_m\} \in \Re^n$ (véase sección 2.3.).

3.1. Naïve Bayes

El clasificador Naïve Bayes (NB) es un algoritmo de aprendizaje probabilístico que deriva de la teoría de decisión bayesiana (Mitchell, 1997). La probabilidad de un documento d en la clase c denotado por $P(c|d)$ se calcula como:

$$\mathrm{P(c|d)} \propto P(c) \prod_{k=1}^{m} P(t_k|\mathrm{c}) \qquad (\mathbf{8})$$

donde $P(t_k|\,\mathrm{c})$ es la probabilidad condicional de que la característica t_k ocurra en un documento de clase c y $P(c)$ es la probabilidad *a priori* de que un documento ocurre en la clase c. $P(t_k|\,\mathrm{c})$ puede ser usado para medir cuánta evidencia t_k aporta respecto a que c sea la clase correcta (Manning et al., 2008). En la clasificación de documento, la clase a la que pertenece se determina encontrando la clase más probable o máxima *a posteriori* (MAP), c_{MAP}, definida por:

$$c_{MAP} = \underset{c \in c_k}{\operatorname{argmax}}\, P(c|d) = \underset{c \in c_k}{\operatorname{argmax}}\, P(c) \prod_{k=1}^{m} P(t_k|\mathrm{c}) \qquad (\mathbf{9})$$

La ecuación (**9**) implica la multiplicación de muchas probabilidades condicionales, una para cada característica. En la práctica, la multiplicación de probabilidades a menudo se convierte en una suma de logaritmos de probabilidades y, por lo tanto, la maximización de la ecuación es realizada alternativamente mediante la siguiente expresión:

$$c_{MAP} = \underset{c \in c_k}{\operatorname{argmax}} \left[\log P(c) + \sum_{k=1}^{m} \log P(t_k|\mathrm{c}) \right] \quad (\mathbf{10})$$

Todos los parámetros del modelo, es decir, las clases *a priori* y las distribuciones de probabilidad de las características, pueden estimarse con frecuencias relativas desde el conjunto de entrenamiento d. Note que cuando una clase y una característica de documento no se producen juntas en el conjunto de entrenamiento, la estimación de probabilidad basada en la frecuencia correspondiente será cero, lo que haría que el lado derecho de la ecuación (**10**) quede indefinido. Este problema se puede mitigar incorporando algunas correcciones como el suavizado de Laplace en todas las estimaciones de probabilidad.

NB es un modelo de aprendizaje de probabilidad simple que puede ser implementado de forma eficiente con una complejidad lineal. Se aplica un supuesto *simplista* o *ingenuo* de que la presencia o ausencia de una característica en una clase es completamente independiente de cualquier otra característica. A pesar del hecho de que esta suposición demasiado simplificada es a menudo imprecisa (en particular para problemas de dominio de textos), NB es uno de los clasificadores más ampliamente utilizados y posee varias propiedades que lo hacen sorprendentemente útil y preciso (Berry & Kogan, 2010).

3.2. Support Vector Machine

Support Vector Machine (SVM) (Berry & Kogan, 2010) ha sido considerado como el algoritmo más prometedor en la clasificación de textos. Los SVM son clasificadores lineales que funcionan en un espacio de características de alta dimensión, que es un mapeo no lineal del espacio de entrada del problema en cuestión. En el espacio transformado, un SVM construye un hiperplano de separación que maximiza la distancia entre las muestras de entrenamiento de dos clases. Esto se hace seleccionando dos hiperplanos paralelos que son tangentes al menos a una muestra de su clase; dichas muestras en los hiperplanos tangenciales se denominan vectores de soporte. La distancia entre los dos planos tangenciales es el margen del clasificador, que debe ser maximizado, y es por eso por lo que un SVM lineal también

es conocido como un clasificador máximo de margen. Una ventaja de trabajar en un espacio de características de alta dimensión es que, en muchos problemas, la tarea de clasificación no lineal en el espacio de entrada original se convierte en una tarea de clasificación lineal en el espacio de características de alta dimensión. SVM trabaja en el espacio de características de alta dimensión sin incorporar ninguna complejidad computacional adicional.

La fortaleza de los SVM proviene de dos propiedades importantes que poseen: representación del *kernel* y optimización de márgenes. En los SVM, la asignación a un espacio de características de alta dimensión y el aprendizaje de la tarea de clasificación en ese espacio sin ninguna complejidad computacional adicional se logran mediante el uso de una función de *kernel*. Una función de *kernel* puede representar el producto punto de las proyecciones de dos puntos de datos en un espacio de características de alta dimensión. El espacio de alta dimensión utilizado depende de la selección de una función específica del *kernel*. La función de clasificación utilizada en las SVM se puede escribir en términos de los productos puntos de los puntos de datos de entrada. Por lo tanto, utilizando una función de *kernel*, la función de clasificación se puede expresar en términos de productos puntos de proyecciones de puntos de datos de entrada en un espacio de características de alta dimensión. Con las funciones del *kernel* no se realiza una asignación explícita de puntos de datos al espacio de dimensión superior, sino que les da a los SVM la ventaja de aprender la tarea de clasificación en ese espacio de dimensión superior.

La segunda propiedad de los SVM es la forma en que se llega a la mejor función de clasificación. Las SVM minimizan el riesgo de un sobreajuste de los datos de entrenamiento al determinar la función de clasificación (un hiperplano) con un margen de separación máximo entre las dos clases. Esta propiedad proporciona a las SVM una muy poderosa capacidad de generalización en la clasificación.

Esto es aplicable solo para tareas de clasificación binaria, por lo que para usar este método para la clasificación de textos (problema de clases múltiples) tiene que ser tratado como una serie de problemas de clasificación binaria.

En las SVM, la función de clasificación es un hiperplano que separa las diferentes clases de datos.

$$\langle \boldsymbol{w}, x \rangle + b = 0 \qquad \textbf{(11)}$$

La notación $\langle \boldsymbol{w}, x \rangle$ representa el producto punto entre los coeficientes del vector normal $\boldsymbol{w}$, que es perpendicular al hiperplano, y el vector de variables x. El escalar b es un término de sesgo.

La solución al problema de clasificación es especificada entonces por el vector normal $\boldsymbol{w}$. Es posible demostrar que el vector $\boldsymbol{w}$ puede ser escrito como una combinación lineal de los puntos de datos x_i, con $i = 1, \ldots, m$, es decir, $\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i x_i \,, \alpha_i \geq 0$. Los puntos de datos x_i con α_i no ceros son llamados vectores de soporte.

Una función *kernel* k puede ser definida como $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$, donde $\Phi: X \rightarrow H$ es un mapeo de puntos en el espacio de entrada X hacia un espacio hiper-dimensional H. Como se puede apreciar, la función de *kernel* mapea implícitamente los puntos de datos de entrada en un espacio de dimensión superior y devuelve el producto de puntos sin realizar el mapeo o calcular el producto de puntos. Hay varias funciones de *kernel* sugeridas para SVM. Algunas de las funciones *kernel* ampliamente usadas en la literatura incluyen función lineal, $k(x_1, x_2) = \langle x_1, x_2 \rangle$; función base radial Gaussiana (RBF), $k(x_1, x_2) = e^{-\sigma \|x_1 - x_2\|^2}$, y función polinomial, $k(x_1, x_2) = \langle x_1, x_2 \rangle^d$. La selección de una función de *kernel* específica para una aplicación depende de la naturaleza de la tarea de clasificación y del conjunto de datos de entrada. Como se puede inferir, el rendimiento de los SVM depende en gran medida de la función de *kernel* específica utilizada. Según Joachims (1998), en la tarea de clasificación de textos, el *kernel* lineal muestra un desempeño comparable a alternativas no lineales.

La función de clasificación en (**11**) tiene una representación dual de la siguiente manera, donde y_i son las etiquetas de las clases de los puntos de entrada.

$$\sum_i \alpha_i \gamma_i \langle x_i, x \rangle + b = 0 \quad \textbf{(12)}$$

Usando una función *kernel* k, la función de clasificación dual arriba en el espacio de alta dimensión H puede ser escrita como

$$\sum_i \alpha_i \gamma_i k \langle x_i, x \rangle + b = 0 \quad \textbf{(13)}$$

Como se mencionó anteriormente, en los SVM la mejor función de clasificación es el hiperplano que tiene el margen máximo que separa las clases. El problema de encontrar el hiperplano del margen máximo se puede formular como un problema de programación cuadrática. Con la representación dual de la función de clasificación anterior en el espacio de alta dimensión H, los coeficientes α_i de la mejor función de clasificación se encuentran resolviendo el siguiente problema de programación cuadrática (dual).

$$\max_w W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j \gamma_i \gamma_j k(x_i, x_j) \quad \textbf{(14)}$$

$$sujeto\ a\ 0 \leq \alpha_i \leq \frac{C}{m}\ (i = 1, \ldots, m);\ \sum_{i=1}^{m} \alpha_i \gamma_i = 0$$

El parámetro C en la formulación anterior es llamado *parámetro de costo* del problema de clasificación. El parámetro de costo representa el valor de penalización utilizado en los SVM para clasificar un punto de datos de entrada. Un valor alto de C dará como resultado una función de clasificación compleja con una clasificación errónea mínima de los datos de entrada, mientras que un valor bajo de C produce

una función de clasificación que es más simple. Por lo tanto, establecer un valor apropiado para C es crítico para el rendimiento de los SVM.

El problema de optimización descrito arriba es muy desafiante cuando el conjunto de datos es muy grande, tal que la complejidad computacional asciende al cuadrado del tamaño del conjunto de datos. Las complejidades computacionales y de almacenamiento pueden ser reducidas dividiendo el conjunto de datos de entrenamiento hacia un número de *chunks* (particiones) y extrayendo vectores de soporte de cada uno de ellos. Los vectores de soporte pueden posteriormente ser combinados.

El mismo procedimiento puede ser utilizado incorporando nuevos documentos hacia el conjunto existente de vectores de soporte. Se puede demostrar que la solución final es tan buena como si todos los documentos fueran procesados juntos (Aas & Eikvil, 1999).

3.3. Random Forests

El algoritmo Random Forests (Breiman, 2001) es una técnica de ensamble de árboles de decisión, en cuya implementación *bagging* cada árbol en un conjunto de árboles de decisión se construye a partir de una muestra *bootstrap* de vectores de características de la data de entrenamiento. Cada muestra *bootstrap* de vectores de características se obtiene a través de un muestreo aleatorio repetido con reemplazo hasta que el tamaño de la muestra *bootstrap* coincide con el tamaño del subconjunto de entrenamiento original. Esto ayuda a reducir la varianza del clasificador, reduciendo la posibilidad de sobreajuste con la muestra de entrenamiento. Cuando se construye cada árbol de decisión, solo se considera un subconjunto de las n características seleccionadas al azar para construir cada nodo de decisión, lo cual evita correlaciones entre los árboles.

El algoritmo Random Forests se usa para realizar la clasificación de documentos y funciona de la siguiente manera:

Paso 1: muestra aleatoria de "m" registros con el que se entrena cada árbol, con $m < M$, donde M es el número para la muestra completa. Aproximadamente se selecciona 63,2% de los datos de entrenamiento usando el método de *bootstrap* (Liu et al., 2015).

Paso 2: construir un árbol de decisión con la muestra extraída, el cual no se poda. Para la construcción del árbol se usan, por defecto, $\sqrt{n}$ características entre las n que se disponen (problema de clasificación).

Paso 3: repetir los pasos 1 y 2. Construir un gran número de árboles de decisión y desarrollar la secuencia de clasificación de árboles de decisión $\{h_1(X), h_2(X), \ldots, h_{ntree}(X)\}$.

Durante el modelamiento de Random Forest, los datos provienen de una muestra *bootstrap*, de modo que aproximadamente 36,8% de las muestras, llamadas Out-Of-Bag (OOB), no han sido extraídas. Los datos de esta muestra son usados como un conjunto de datos de prueba para examinar el rendimiento del modelo a través de la tasa de error OOB estimada. Breiman (2001) demostró que esta es insesgada, y hace que el modelo de Random Forests no parezca sobrestimado.

Paso 4: la clasificación final se determina por cada voto registrado a partir de los resultados de la clasificación del árbol de decisión.

Esto se puede expresar de la siguiente manera: h_i es un modelo de árbol de decisión individual, Y representa la variable de salida (o target) e I( ) es una función indicadora.

$$H(x) = \underset{Y}{\operatorname{argmax}} \sum_{i=1}^{n} I(h_i(x) = Y) \qquad (15)$$

Cada árbol entrega una clasificación sobre los datos sobrantes (OOB), y decimos el árbol "votos" para esa clase. El bosque elige la clasificación que tiene más votos sobre todos los árboles del bosque. Esta es la puntuación de Random Forest y el porcentaje de votos recibidos por una clase es la probabilidad pronosticada. Por ejemplo, en un modelo con respuesta binaria dado, si se consideran 500 árboles, y un caso es OOB en 200 de ellos, donde 160 votan por la clase 1 y 40 votan por la clase 2, el modelo de Random Forest lo clasifica como clase 1, donde la probabilidad asociada a ese caso es de 0.80 (160/200).

De las n características seleccionadas aleatoriamente para construir cada uno de los nodos de decisión, se selecciona la condición respecto de la clase c que reduce mejor la métrica de impurezas Gini g de los datos, la cual indica cuán a menudo un elemento seleccionado aleatoriamente del conjunto sería etiquetado incorrectamente si fuese etiquetado de manera aleatoria de acuerdo con la distribución de las etiquetas en el subconjunto. La impureza de Gini alcanza su mínimo (cero) cuando todos los casos de un nodo corresponden a una sola categoría de destino y entre mayor sea su valor, el clasificador es más incierto acerca de si un vector de características pertenece a una clase o a otra. El cálculo de la impureza de Gini para un nodo del árbol se obtiene a través de la siguiente expresión:

$$g = 1 - \sum_{c=1}^{m} P_c^2 \qquad \textbf{(16)}$$

Donde P_c es la probabilidad de que un documento quede etiquetado en la clase c.

4. Métricas de desempeño

Un problema muy importante en la clasificación de textos es cómo evaluar el desempeño de los clasificadores (métodos o modelos). Muchas medidas han sido usadas, cada una de las cuales ha sido diseñada para evaluar algún aspecto del desempeño de clasificación de un sistema. En esta sección se describen algunas de las métricas que han sido reportadas en la literatura.

Un enfoque común para el problema de clasificación con clases múltiples es llevar la tarea hacia problemas disjuntos de clasificación binaria. Para cada clase y cada documento uno determina si el documento pertenece a la clase de interés (clase positiva) o no (clase negativa). Cuando se evalúa el desempeño de un clasificador, cuatro cantidades que se desprenden de la matriz de confusión son de interés para cada clase (Cuadro 1).

Cuadro 1. Matriz de Confusión para un problema de dos clases

	Predicho positivo	Predicho negativo
Observado positivo	VP	FN
Observado negativo	FP	VN

Fuente: Elaboración propia.

Donde:

VP: Número de documentos correctamente predichos para la clase de interés.
FP: Número de documentos incorrectamente predichos para la clase de interés.
FN: Número de documentos incorrectamente rechazados para la clase de interés.
VN: Número de documentos correctamente rechazados para la clase de interés.

A partir de estas cantidades, se definen las siguientes métricas de desempeño:

- *Recall*: mide la precisión del clasificador (modelo) sobre los casos de la clase de interés.

$$recall = \frac{\text{VP}}{VP + FN} \tag{17}$$

- *Precision*: mide la precisión del clasificador (modelo) sobre los casos predichos de la clase de interés.

$$precision = \frac{\text{VP}}{VP + FP} \tag{18}$$

- *Accuracy*: mide la precisión global de las predicciones realizadas por el clasificador (modelo).

$$accuracy = \frac{\text{VP} + \text{VN}}{VP + FN + FP + VN} \tag{19}$$

Micro & macro promedios: para evaluar el desempeño promedio a través de las clases existen dos métodos convencionales (Aas & Eikvil, 1999), a saber: macro – promedio y micro – promedio. Para el macro – promedio el puntaje se determina mediante el cálculo de las métricas de desempeño por clase y luego promediando estas para calcular las medias globales. Para el micro – promedio el puntaje se determina calculando primero los totales VP, FP, FN y VN para todas las clases y luego usando estos totales para calcular las métricas de desempeño. Hay una importante distinción entre los dos tipos de promedios. En el micro – promedio se le

da igual peso a cada documento, mientras que en el macro – promedio se otorga igual peso a cada clase.

Punto de equilibrio: Las métricas de desempeño anteriores pueden resultar engañosas cuando se examinan por sí solas. Usualmente un clasificador exhibe un *trade-off* entre $recall$ y $precision$, donde obtener un alto $recall$ típicamente significa sacrificar $precision$ y viceversa. Si $recall$ y $precision$ son configurados para tener igual valor, entonces este valor es llamado *punto de equilibrio* del sistema. El *punto de equilibrio* ha sido comúnmente utilizado en evaluaciones de clasificación de textos (Aas & Eikvil, 1999).

Medida F: otro criterio de evaluación que combina $recall$ y $precision$ es la medida F:

$$F_\beta = \frac{(1+\beta^2)\cdot\ recall\ \cdot precision}{\beta^2 \cdot precision + recall} \qquad \textbf{(20)}$$

donde β es un parámetro que permite diferentes ponderaciones de $recall$ y $precision$. Para este trabajo se ha considerado $\beta = 1$, es decir, la medida F_1 corresponde a la media armónica entre el $recall$ y la $precision$.

5. Aplicación a las glosas de la ENE

Tal como se mencionó antes, en este documento se aplicaron los métodos de *text mining* descritos en las secciones anteriores sobre las glosas de la Encuesta Nacional de Empleo (ENE). Esto, para generar una alternativa a la tarea de clasificación de texto que debe ser ejecutada mensualmente por los funcionarios del INE de forma manual.

5.1. El conjunto de datos

El conjunto de datos utilizado en los ejercicios correspondió a 183.784 documentos de la base de datos de 2017 de la ENE, de las glosas que se derivan de las respuestas declaradas por los informantes respecto a su "oficio, labor u ocupación" y del "sector económico" al que pertenece la empresa en la que trabajan, que vienen dadas por las preguntas B1 y B14b del cuestionario.

Las variables objetivo para el entrenamiento de los clasificadores son los de actividad económica del CIUO (Clasificador Internacional Uniforme de Ocupaciones) y del CAENES (Clasificador de Actividades Económicas Nacional para Encuestas Sociodemográficas) a uno y dos dígitos. El clasificador CIUO (DANE, 2005) es una de las principales clasificaciones de las que la Organización Internacional del Trabajo (OIT) es responsable, y es una herramienta que sirve para organizar los empleos en una serie de grupos definidos claramente en función de las tareas que caracterizan a cada empleo. Por su parte, el CAENES (INE, 2016) está basado en el Clasificador Chileno de Actividades Económicas CIIU4.CL 2012, cuya estructura facilita la clasificación de actividades económicas. Esto, dado que la categoría más desagregada, denominada clase, puede agrupar varias subclases, clases o grupos del CIIU4.CL 2012 o abrir otras, según sea necesario, permitiendo que el nivel de detalle requerido sea el idóneo para las encuestas de hogares.

En definitiva, se tienen cuatro variables objetivo para el entrenamiento de los clasificadores, a saber: CIUO-1, CIUO-2, CAENES-1 y CAENES-2. Estas variables dan cuenta de los clasificadores de actividad económica CIUO a uno y dos dígitos, y CAENES a uno y dos dígitos, respectivamente. El CIUO-1 está compuesto por un total de 10 clases, el CIUO-2 lo conforman 27 clases, el CAENES-1 lo integran 21 clases y el CAENES-2 está compuesto por 83 clases.

Es importante mencionar que los documentos con los cuales se realiza el proceso de clasificación de CIUO-1 y CIUO-2 son el resultado de la concatenación de la

"ocupación, descripción de tareas" y "sector económico" al que pertenece la empresa, respondidos por los informantes, mientras que para CAENES-1 y CAENES-2 corresponden al "sector económico" al que pertenece la empresa, declarado por el informante.

En los datos utilizados para los clasificadores CAENES-1 y CAENES-2 se detectó presencia de valores nulos (*NULL*) y perdidos (*missing*) en los documentos. Por lo tanto, fueron removidos, quedando 177.176 documentos en el conjunto de datos. Para la construcción de los clasificadores, los conjuntos de datos de entrada fueron divididos en un conjunto de entrenamiento, compuesto por el 80% de los casos, y en uno de prueba, compuesto por el 20%. En la Tabla 1. se resume la composición de los conjuntos de entrenamiento y prueba para la construcción de cada clasificador.

Tabla 1. Conjuntos de datos de entrenamiento y prueba para CIUO y CAENES

Clasificador	No. entrenamiento	No. prueba
CIUO-1	147.027	36.757
CIUO-2	147.027	36.757
CAENES-1	141.740	35.436
CAENES-2	141.740	35.436

Fuente: Elaboración propia.

5.2. Preparación de los datos

Los cuerpos de todos los documentos fueron convertidos desde el formato original (i.e. cadenas de caracteres) a vector de palabras. En lo que sigue, se describen los pasos de este procedimiento:

1. Las palabras individuales fueron extraídas a través de la normalización y tokenización (véase sección 2.1). Luego, los *stopwords* fueron removidos usando una lista compuesta por 344 palabras frecuentes del español (por ejemplo: ante, de, cual, entonces). Posteriormente, un proceso de *stemming* fue llevado a cabo mediante la aplicación de la librería de R *hunspell* (Ooms, 2017). Este procedimiento resultó en 37.685 palabras únicas.

2. La Ganancia de Información (IG) fue utilizada como métrica para la selección de características, lo que permite discriminar qué palabras aportan una mayor información para la clasificación. Los resultados de esta métrica pueden verse en el **anexo 1**. Además, se eliminaron palabras bajo un umbral de frecuencia

definido en todos los documentos de la colección (véase sección 2.2.), que fue de cinco. De esta manera, para los clasificadores CIUO quedaron finalmente 8.448 palabras y, en consecuencia, la reducción de dimensionalidad entregó como resultado una matriz de documento – término de 147.027 *x* 8.448 para el conjunto de datos de entrenamiento. Por su parte, para los clasificadores CAENES quedaron 5.346 palabras, lo que entregó como resultado una matriz de documento – término de 141.740 *x* 5.346 para el conjunto de datos de entrenamiento.

3. La medida de pesos "*tf – idf*" (véase sección 2.3.) fue usada para la indexación de palabras en la matriz de documento – término.

5.3. Utilización de los métodos de *machine learning*

Este trabajo buscó aplicar y evaluar tres técnicas de *machine learning* para la clasificación de documentos (glosas) en algunos de los grupos de actividad económica definidos por el CIUO (uno y dos dígitos) y el CAENES (uno y dos dígitos). La clasificación se realizó en modo supervisado, es decir, el entrenamiento se llevó a cabo con ejemplos de clases (grupos de actividad) previamente etiquetados.

Los parámetros para las técnicas que se usaron en la evaluación comparativa se establecieron a partir de lo obtenido en las pruebas en el conjunto de datos. No se hizo ningún ajuste adicional de los parámetros. Si bien la puesta a punto de los parámetros a conjuntos de datos específicos puede ser beneficiosa, la consideración de configuraciones generalmente aceptadas es más típica en la práctica. La necesidad de un esfuerzo y tiempo significativos para el ajuste fino de los parámetros a menudo puede ser un impedimento para el uso práctico, y también puede conducir a problemas de sobreajuste de datos específicos.

Para Naïve Bayes (NB), las clases *a priori* y las distribuciones de probabilidad de las características fueron estimadas a partir de las frecuencias desde el conjunto de datos de entrenamiento.

Para Support Vector Machine (SVM) se optó por la utilización de un *kernel* lineal, y tras realizar la optimización del parámetro de coste C a través de una *cross – validation* a 10 *folds,* se obtuvo un valor óptimo de 0,75 para CIUO y CAENES a un dígito, y un valor óptimo de 1,0 para CIUO y CAENES a dos dígitos.

Para Random Forest (RF) se estableció el número de características en $\sqrt{n}$ y un número de árboles $T = 1000$.

La implementación de estos tres métodos fue efectuada mediante el *software* estadístico R (R Core Team, 2018).

5.4. Resultados

A continuación se muestran los resultados obtenidos de la aplicación de los métodos antes descritos sobre las glosas de la ENE. Para evaluar el desempeño de los distintos métodos se utilizó la *precision*, el *recall* y la medida F_1 (véase sección 4).

5.4.1. CIUO – 1

En la tabla 2. se muestran en orden descendente los números de documentos en la colección utilizados para el entrenamiento y prueba de los modelos para cada una de las clases que componen el CIUO-1, donde destaca la frecuencia de la clase 9, que se refiere al grupo de trabajadores no calificados.

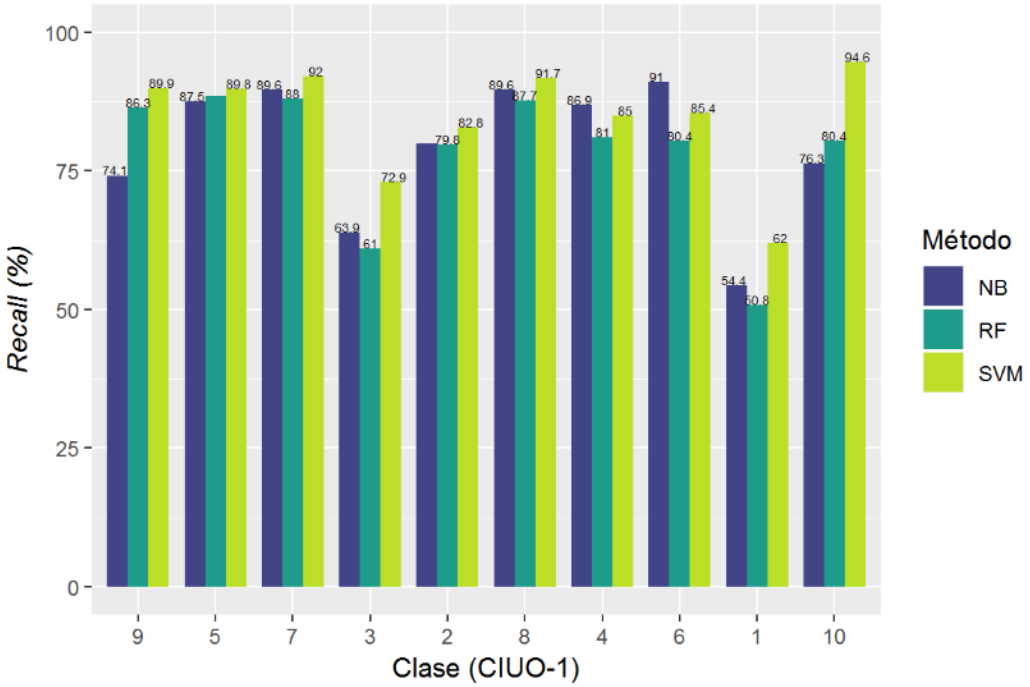
Tabla 3. Conjuntos de datos de entrenamiento y prueba para CIUO – 1

Clase	No. entrenamiento	No. prueba
9. Trabajadores no calificados	35.469	8.788
5. Trabajadores de los servicios y vendedores	22.337	5.606
7. Oficiales, operarios, artesanos y trabajadores de la industria manufacturera, de la construcción y de la minería	20.378	4.984
3. Técnicos, postsecundarios no universitarios y asistentes	16.309	4.005
2. Profesionales universitarios científicos e intelectuales	15.900	4.050
8. Operadores de instalaciones, de máquinas y ensambladores	12.942	3.268
4. Empleados de oficina	12.667	3.228
6. Agricultores, trabajadores y obreros agropecuarios, forestales y pesqueros	7.223	1.816
1. Miembros del Poder Ejecutivo, de los cuerpos legislativos y personal directivo de la administración pública y de empresas	2.616	688
10. Fuerza pública	1.186	324

Fuente: Elaboración propia.

En el gráfico 1. se observa que los tres métodos tienen un desempeño relativamente estable a través de las clases. SVM es el que exhibe el mejor desempeño en todas las clases, principalmente sobre las clases 9,5 y 7, que son las que concentran el mayor número de documentos en la colección, con un *recall* en torno al 90%. En contraste, las clases 3 y 1 muestran desempeños relativamente bajos para todos los métodos, con un *recall* bajo el 70%.

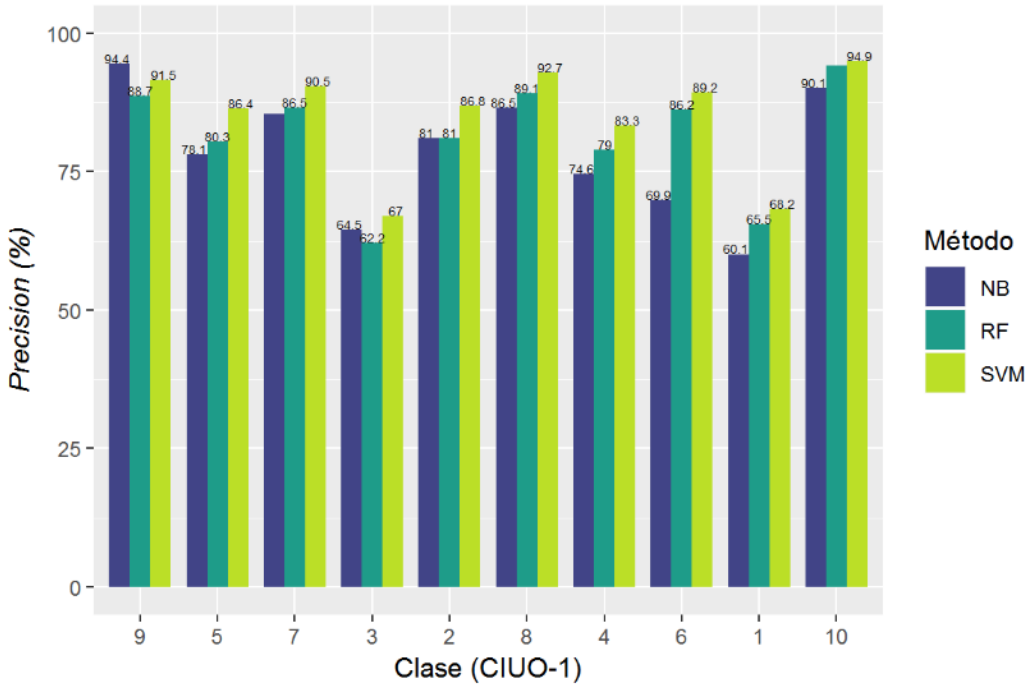
Gráfico 1. Desempeño de *Recall* según método para CIUO – 1



Fuente: Elaboración propia.

En el gráfico 2. se observa que NB funciona muy bien en la predicción sobre la clase 9, que es la más poblada en el conjunto de datos, alcanzando una *precision* del 94%. Los tres métodos muestran un desempeño relativamente estable y sobre el 80%, principalmente sobre las clases con mayor número de documentos en la colección, es decir, 9,5 y 7. Además, tal como en el *recall*, las clases 3 y 1 exhiben un desempeño relativamente bajo, en torno al 68%.

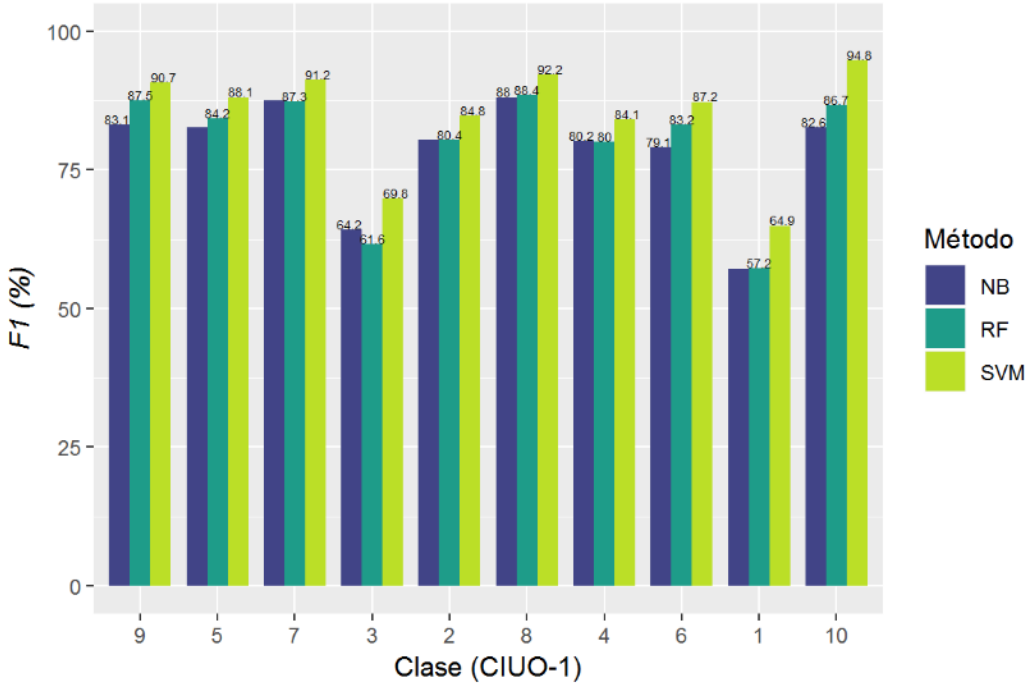
Gráfico 2. Desempeño de *Precision* según método para CIUO – 1



Fuente: Elaboración propia.

En el gráfico 3. se puede observar que la medida F_1 muestra un comportamiento similar al observado en el *recall*, donde SVM muestra el mejor desempeño de los tres métodos.

Gráfico 3. Desempeño de la medida F_1 según método para CIUO – 1



Fuente: Elaboración propia.

5.4.2. CIUO – 2

En la tabla 3. se muestran en orden descendente los números de documentos en la colección utilizados para el entrenamiento y prueba de los modelos para cada una de las clases que componen el CIUO-2, donde destacan las frecuencias de las clases 91 y 52, que se refieren a actividades de ventas y servicios.

Tabla 2. Conjuntos de datos de entrenamiento y prueba para CIUO – 2

Clase	No. entrenamiento	No. prueba
91. Trabajadores no calificados de servicios (excepto el personal doméstico y afines)	19.503	4.794
52. Personal de los servicios de protección y seguridad	12.469	3.156
51. Trabajadores de los servicios personales	9.868	2.450
83. Conductores de vehículos y operadores de equipos pesados móviles	9.840	2.549
34. Otros técnicos, postsecundarios no universitarios y asistentes	9.756	2.354
92. Personal doméstico, aseadores, lavanderos, planchadores y afines	9.128	2.285
71. Oficiales y operarios de la industria extractiva	8.544	2.077
41. Oficinistas	8.501	2.140

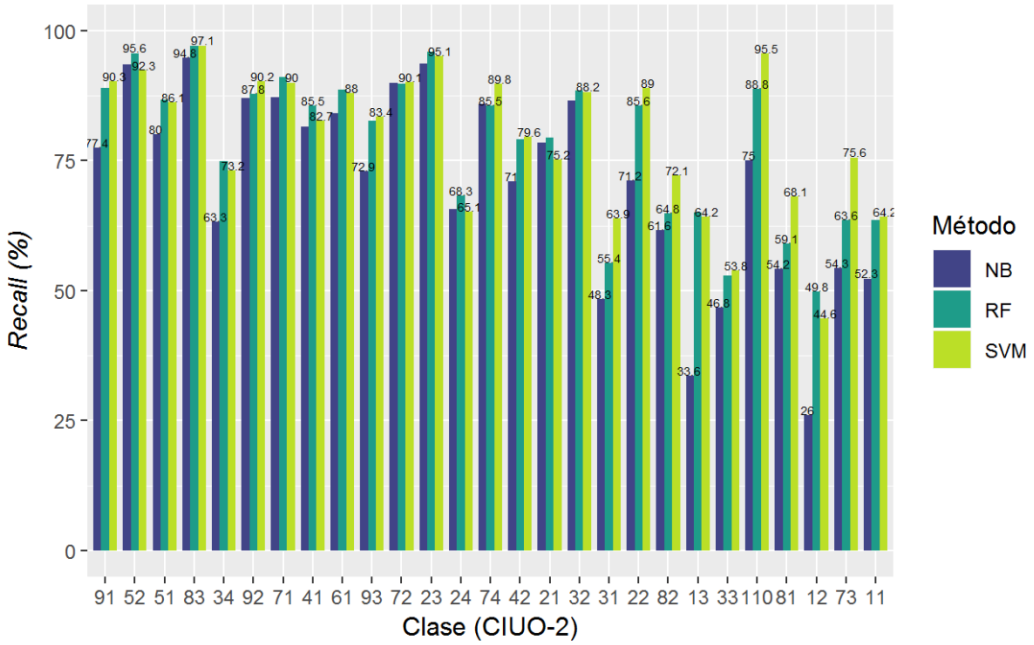
Clase	No. entrenamiento	No. prueba
61. Agricultores y trabajadores forestales, pecuarios y pesqueros	7.223	1.816
93. Obreros de la minería, la construcción, la industria manufacturera y el transporte	6.838	1.709
72. Oficiales y operarios de la construcción	6.024	1.442
23. Profesionales de la educación	5.588	1.378
24. Otros profesionales científicos e intelectuales	5.398	1.393
74. Mecánicos y ajustadores de máquinas y equipos	4.992	1.267
42. Empleados de trato directo con el público	4.166	1.088
21. Profesionales de las ciencias físicas, químicas, matemáticas y de la ingeniería	2.899	793
32. Técnicos y postsecundarios no universitarios de las ciencias biológicas, la medicina y la salud	2.683	658
31. Técnicos y postsecundarios no universitarios de las ciencias físicas, químicas, la ingeniería y afines	2.231	583
22. Profesionales de las ciencias biológicas, la medicina y la salud	2.015	486
82. Operadores de máquinas y ensambladores	1.978	462
13. Directores de departamentos públicos y privados	1.657	446
33. Asistentes de enseñanza e instructores de educación formal, especial y vocacional	1.639	410
110. Oficiales de las Fuerzas Militares	1.186	324
81. Operadores de instalaciones fijas y afines	1.124	257
12. Directores y gerentes generales de empresas privadas	837	213
73. Operarios de la metalurgia y afines	818	198
11. Miembros del poder ejecutivo, de los cuerpos legislativos y personal directivo de la administración pública	122	29

Fuente: Elaboración propia.

En el gráfico 4. se observa que los tres métodos muestran comportamientos relativamente estables a través de las clases. SVM exhibe el desempeño más alto en casi todas las clases, con un *recall* en torno al 90% en las clases 91 y 52, que contienen la mayor cantidad de documentos en la colección. RF presenta desempeños similares a los de SVM, particularmente en las clases con mayor número de documentos y, finalmente, NB se encuentra a una distancia de SVM y RF en casi todas las clases.

Además, se observa una consistencia con lo observado en CIUO-1, donde clases como 11, 12, 13, 31, 32 y 33 muestran un patrón similar al observado en las clases 1 y 3, cuyo *recall* oscila alrededor del 60%.

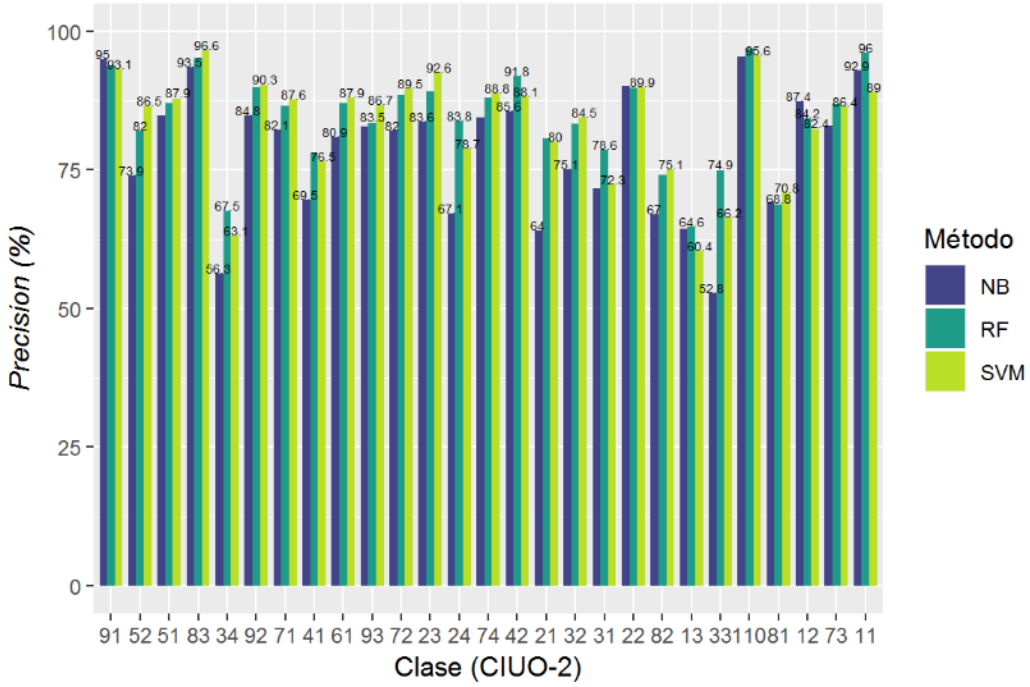
Gráfico 4. Desempeño de *Recall* según método para CIUO – 2



Fuente: Elaboración propia.

En el gráfico 5. se observa que los tres métodos muestran comportamientos relativamente estables a través de las clases. NB exhibe una notable *precision* sobre la clase 91, con un desempeño del 95%. SVM y RF destacan por su paridad en altos desempeños a través de las clases, principalmente en aquellas con un mayor número de documentos, alcanzando una *precision* en torno al 85%.

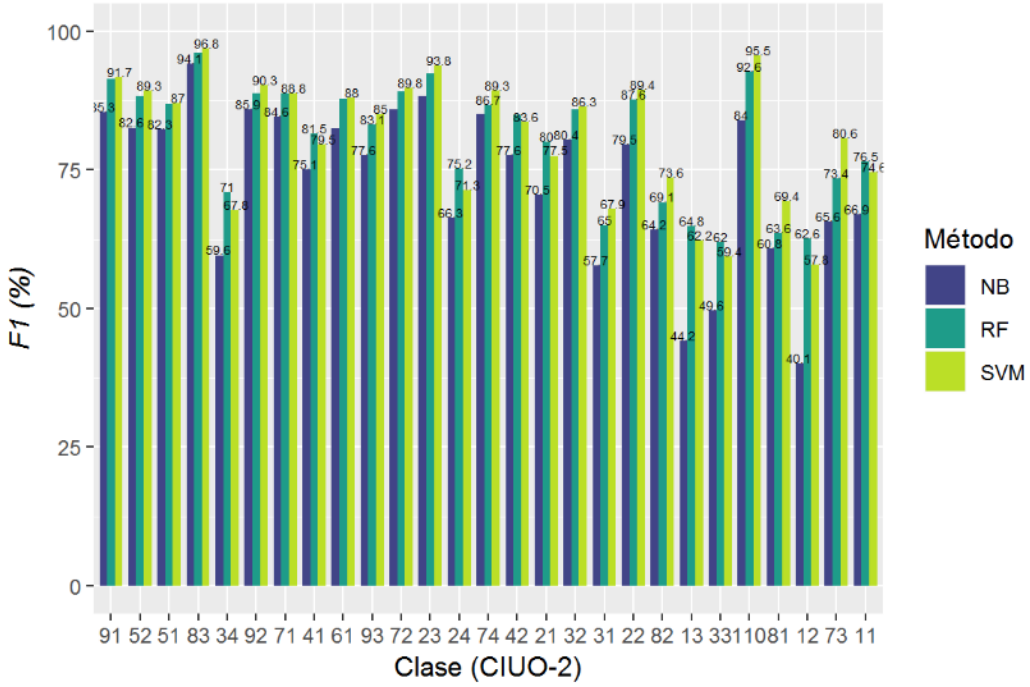
Gráfico 5. Desempeño de *Precision* según método para CIUO – 2



Fuente: Elaboración propia.

El gráfico 6. indica que F_1 muestra un patrón similar al observado en el *recall*, donde SVM presenta los desempeños más altos en casi todas las clases.

Gráfico 6. Desempeño de la medida F_1 según método para CIUO – 2



Fuente: Elaboración propia.

5.4.3. CAENES – 1

En la tabla 4. se muestran en orden descendente los números de documentos en la colección utilizados para el entrenamiento y prueba de los modelos para cada una de las clases que componen el CAENES-1.

Tabla 3. Conjuntos de datos de entrenamiento y prueba para CAENES – 1

Clase	No. entrenamiento	No. prueba
G. Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas	26.718	6.556
A. Agricultura, ganadería, silvicultura y pesca	17.569	4.439
C. Industrias manufactureras	14.688	3.688
P. Enseñanza	13.027	3.201
F. Construcción	11.585	2.840
H. Transporte y almacenamiento	9.179	2.388
O. Administración pública y defensa; planes de seguridad social de afiliación obligatoria	9.064	2.343
Q. Actividades de atención de la salud humana y de asistencia social	8.030	2.048
I. Actividades de alojamiento y de servicio de comidas	6.398	1.575
B. Explotación de minas y canteras	4.707	1.193
S. Otras actividades de servicios	4.258	997
M. Actividades profesionales, científicas y técnicas	3.665	885
N. Actividades de servicios administrativos y de apoyo	3.236	863
K. Actividades financieras y de seguros	2.294	575
J. Información y comunicaciones	1.960	512
R. Actividades artísticas, de entretenimiento y recreativas	1.647	420
L. Actividades inmobiliarias	1.127	288
E. Suministro de agua; evacuación de aguas residuales, gestión de desechos y descontaminación	975	230
D. Suministro de electricidad, gas, vapor y aire acondicionado	806	197
T. Actividades de los hogares como empleadores; actividades no diferenciadas de los hogares como productores de bienes y servicios para uso propio	769	187

Clase	No. entrenamiento	No. prueba
U. Actividades de organizaciones y órganos extraterritoriales	24	5

Fuente: Elaboración propia.

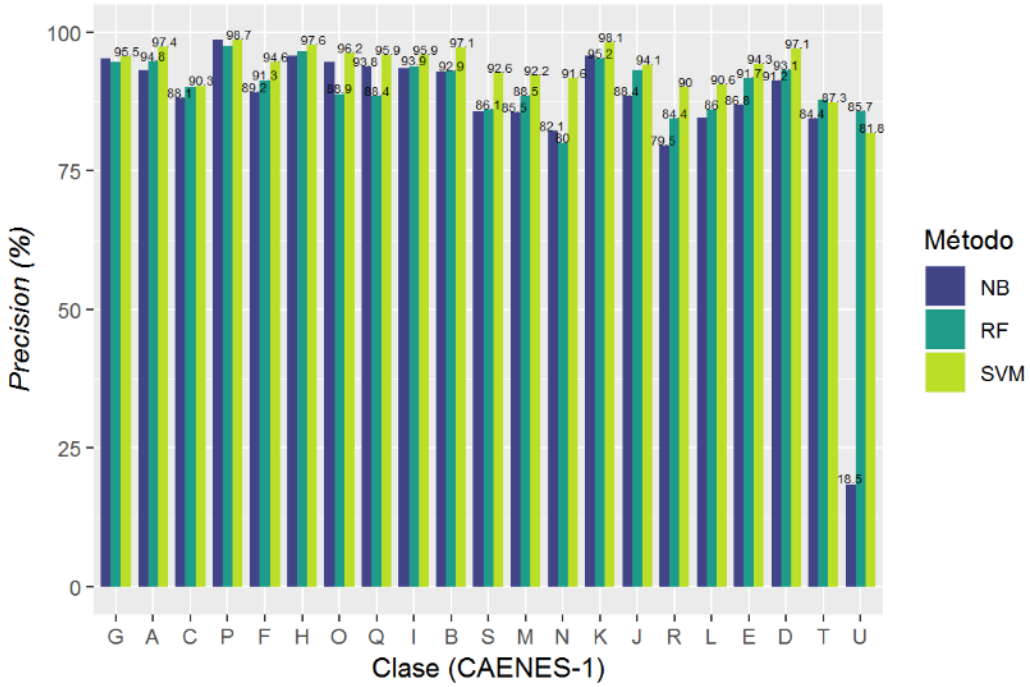
En el gráfico 7. se observa que los tres métodos presentan desempeños muy estables, con un *recall* sobre el 90% en casi todas las clases, lo que es muy satisfactorio. En contraste, se observa en la clase U, que se refiere a actividades de organizaciones y órganos extraterritoriales, un desempeño bajo el 70%; sin embargo, este resultado se debe manejar solo como una referencia descriptiva, dado el bajo número de documentos de entrenamiento (y de prueba) que posee. En general, SVM muestra una leve ventaja en rendimiento respecto a NB y RF.

Gráfico 7. Desempeño de *Recall* según método para CAENES – 1

Fuente: Elaboración propia.

En el gráfico 8. se aprecia que los tres métodos presentan una alta *precision*, donde SVM exhibe el mejor desempeño, alcanzando uno sobre el 90% en casi todas las clases. Tal como en el *recall*, los desempeños observados en U solo merecen una consideración de carácter descriptivo.

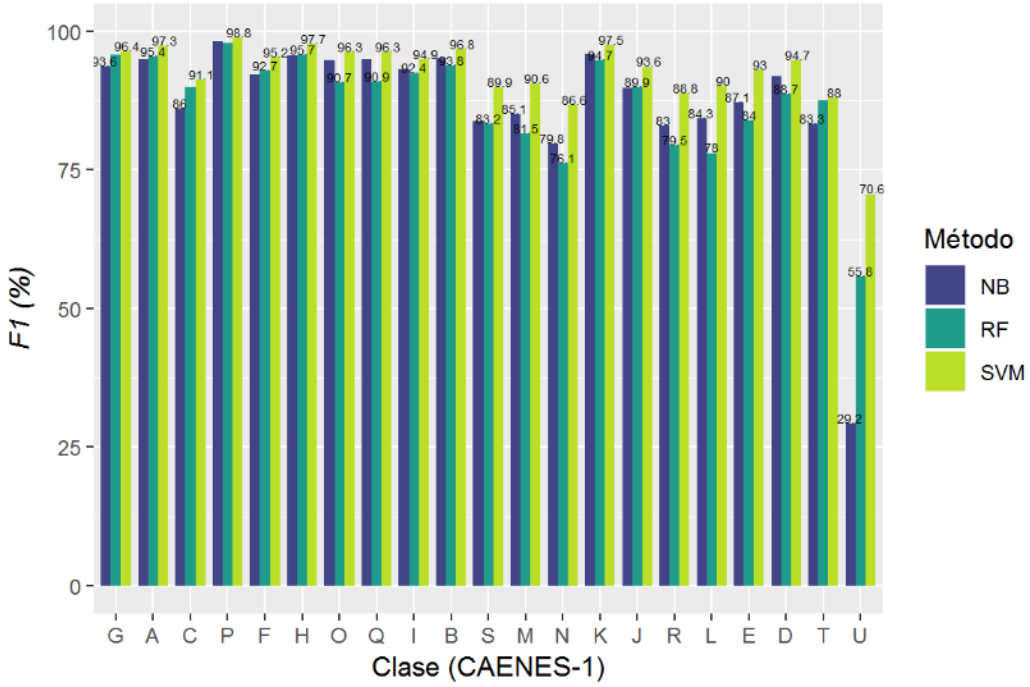
Gráfico 8. Desempeño de *Precision* según método para CAENES – 1



Fuente: Elaboración propia.

En el gráfico 9. se aprecia que la medida F_1 mantiene un patrón similar al observado en la *precision*, donde SVM muestra el mejor desempeño en todas las clases, alcanzando un desempeño sobre el 90% en casi todas las clases.

Gráfico 9. Desempeño de la medida F_1 según método para CAENES – 1



Fuente: Elaboración propia.

5.4.4. CAENES – 2

En la tabla 5. se muestran en orden descendente los números de documentos en la colección utilizados para el entrenamiento y prueba de los modelos para cada una de las clases que componen el top 27 del número de documentos para CAENES-2. Estas clases representan aproximadamente el 86% de los documentos en la colección. En la tabla 5. destaca la frecuencia de documentos que pertenecen a la clase 48, que corresponde al comercio al por mayor y menor, excepto de vehículos automotores y bicicletas. En los gráficos 10., 11. y 12. se muestran los rendimientos para estas clases respecto el *recall*, *precision* y medida F_1, respectivamente. Para revisar el desempeño de las clases restantes, véase el **anexo 2**.

Tabla 4. Conjuntos de datos de entrenamiento y prueba para el Top 27 de CAENES – 2

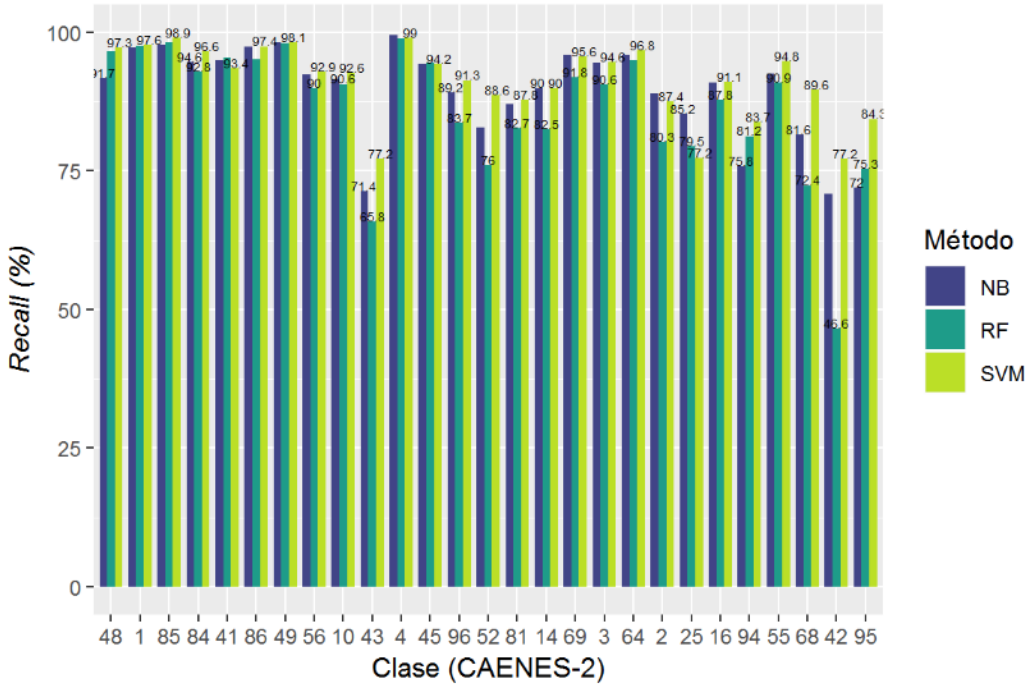
Clase	No. entrenamiento	No. prueba
48. Comercio al por mayor y al por menor, excepto de vehículos automotores y motocicletas	23.908	5.848
1. Agricultura, ganadería, caza y actividades de servicios conexas	14.653	3.684
85. Enseñanza	13.027	3.201
84. Administración pública y de defensa; planes de seguridad social de afiliación obligatoria	9.064	2.343

Clase	No. entrenamiento	No. prueba
41. Construcción de edificios	6.775	1.727
86. Actividades de atención de la salud humana	6.657	1.715
49. Transporte por vía terrestre y transporte por tuberías	6.652	1.703
56. Actividades de servicio de comidas y bebidas	5.221	1.280
10. Elaboración de productos alimenticios	5.071	1.202
43. Actividades especializadas de construcción	3.796	874
4. Extracción y procesamiento de cobre	3.743	977
45. Comercio y reparación de vehículos automotores y motocicletas	2.810	708
96. Otras actividades de servicios personales	2.077	468
52. Almacenamiento y actividades de apoyo al transporte	1.740	466
81. Actividades de servicios a edificios y de paisajismo	1.663	411
14. Fabricación de productos textiles, fabricación de prendas de vestir, productos de cuero y productos conexos	1.624	426
69. Actividades jurídicas y de contabilidad	1.610	384
3. Pesca, acuicultura y actividades de servicios conexas	1.528	399
64. Actividades de servicios financieros, excepto las de seguros y fondos de pensiones	1.486	373
2. Silvicultura, extracción de madera y actividades de servicios conexas	1.388	356
25. Fabricación de productos elaborados de metal y servicios de trabajo de metales, excepto máquinas y equipos	1.301	316
16. Producción de madera y fabricación de productos de madera y corcho, excepto muebles; fabricación de artículos de paja y de materiales trenzados	1.256	342
94. Otras actividades de servicios personales	1.216	288
55. Actividades de servicio de comidas y bebidas	1.177	295
68. Actividades inmobiliarias	1.127	288
42. Comercio y reparación de vehículos automotores y motocicletas	1.014	239
95. Reparación de computadores y de efectos personales y enseres domésticos	965	241

Fuente: Elaboración propia.

En el gráfico 10. se aprecia que los tres métodos presentan un alto desempeño en casi todas las clases, alcanzando un *recall* en torno al 90%, que es bastante satisfactorio. Esto se observa principalmente en las clases con mayor número de documentos de entrenamiento (y de prueba).

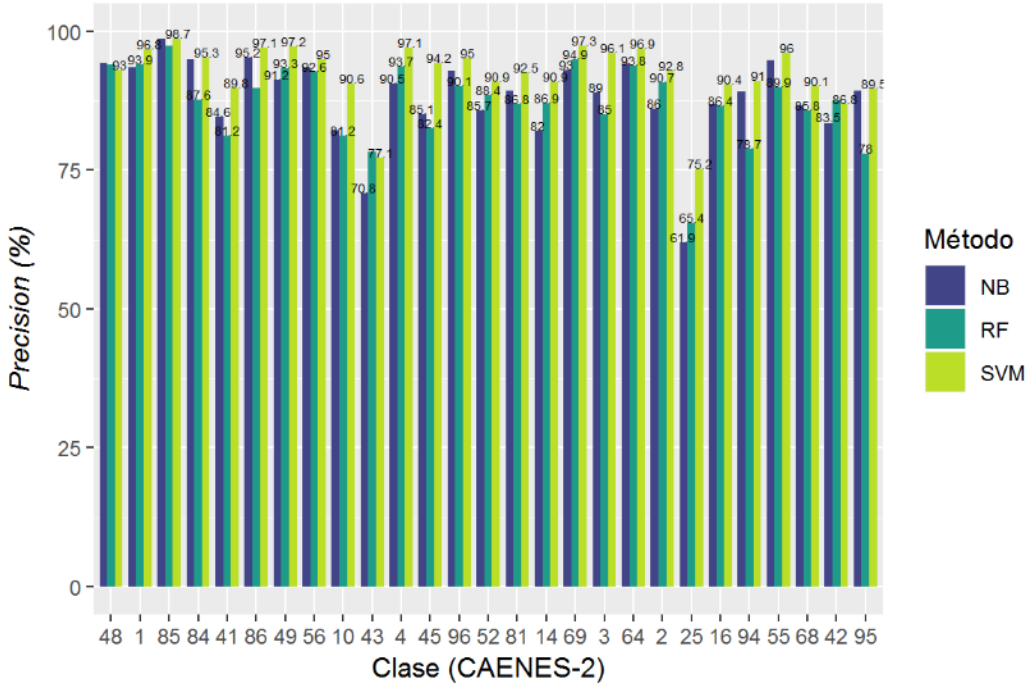
Gráfico 10. Desempeño de *Recall* según método para CAENES – 2. Clases en las primeras 27 posiciones según número de documentos en la colección



Fuente: Elaboración propia.

El gráfico 11. indica que los tres métodos exhiben desempeños relativamente estables con una *precision* en torno al 90%, donde SVM presenta el mejor desempeño en casi todas las clases.

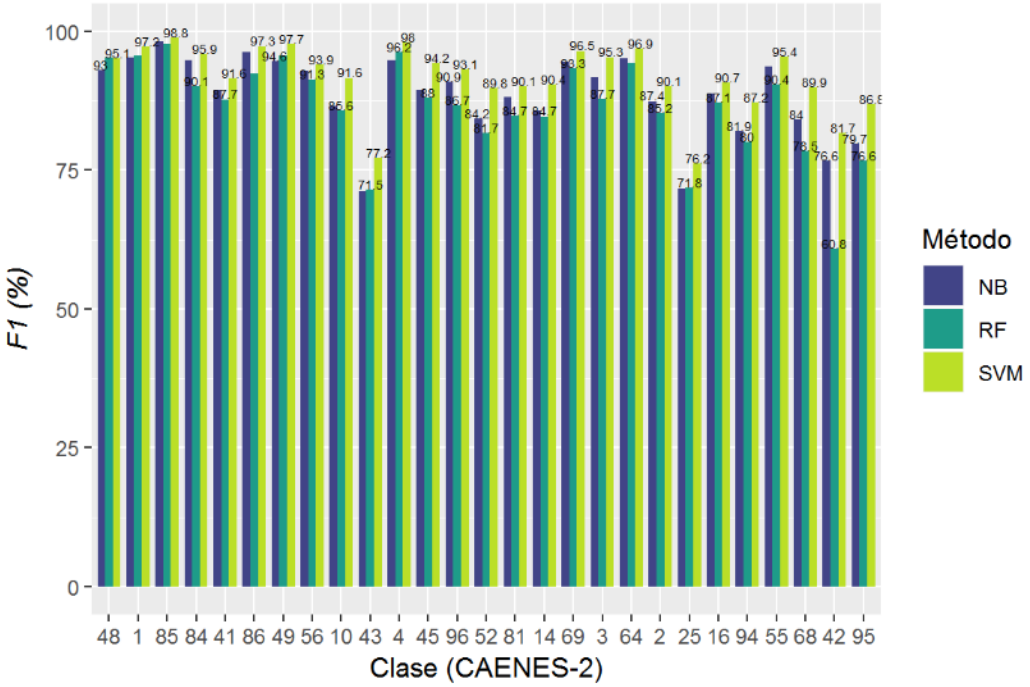
Gráfico 11. Desempeño de *Precision* según método para CAENES – 2. Clases en las primeras 27 posiciones según número de documentos en la colección



Fuente: Elaboración propia.

El gráfico 12. muestra que la medida F_1 presenta un patrón similar a la *precision* y que SVM tiene el mejor desempeño de los tres métodos.

Gráfico 12. Desempeño de la medida F_1 según método para CAENES – 2. Clases en las primeras 27 posiciones según número de documentos en la colección

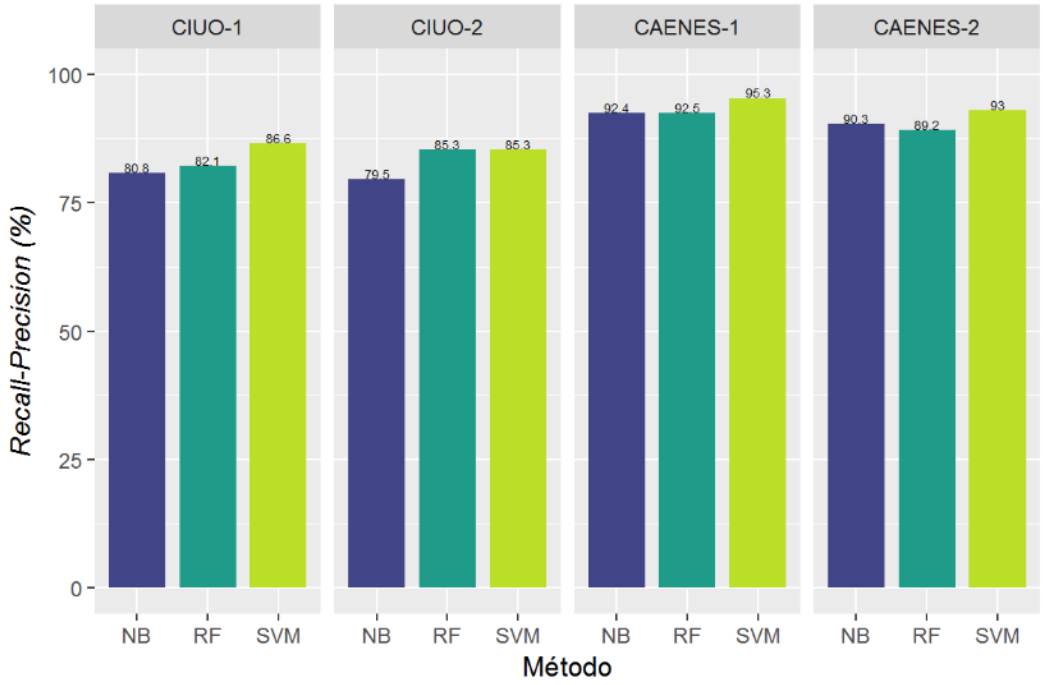


Fuente: Elaboración propia.

5.4.5. Desempeño global

Finalmente, para evaluar la efectividad global de los métodos se utilizó, al igual que en otros trabajos reportados en la literatura, el punto de equilibrio *recall/precision*, que da cuenta del *trade-off* entre *recall* y *precision*. El desempeño promedio de las clases se calculó mediante micro – promedios (véase sección 4.), medida que otorga igual peso a cada documento en la colección. El gráfico 13. muestra el *recall/precision* obtenido tanto para el clasificador CIUO como para el CAENES, considerando todas las clases en cada caso. Se aprecia que los tres métodos son muy competitivos y funcionan bien en la tarea de clasificación, con un *recall/precision* sobre el 80% para el CIUO y sobre el 90% para el CAENES, lo que es bastante satisfactorio. SVM alcanza los más altos desempeños en CIUO-1, CAENES-1 y CAENES-2, con un *recall/precision* de 86,6%, 95,3% y 93,0%, respectivamente. Para CIUO-2, SVM y RF alcanzan el mejor desempeño, con un *recall/precision* de 85,3%.

Gráfico 13. Desempeño de *Recall-Precision* según método para clasificadores CIUO y CAENES



Fuente: Elaboración propia.

5.5. Discusión y comparación de los métodos

Comparada con la clasificación de textos en general, la clasificación de glosas de la ENE representa una especial y muy desafiante tarea de clasificación. Para las diferentes variables objetivo hay distinto número de clases posibles. Para CIUO-1 se

tiene 10 clases, para CIUO-2 se tiene 27, para CAENES-1 se tiene 21 y para CAENES-2 se tiene 83 clases posibles. Aunque algunas de las clases poseen una apariencia similar, aún existen importantes y diferentes características en las clases de actividad económica que no deberían ser pasadas por alto. Por ejemplo, para CIUO-1, en contraste con la clase 10, la clase 9 tiene un vocabulario mucho más extenso. Idealmente, un algoritmo de *machine learning* exitoso utilizado en este dominio de clasificación particular debería usar completamente las posibles diferencias entre las clases de actividad económica y, lo que es más importante, debería ser capaz de perfilar las clases con precisión y conducir solo a un pequeño número de errores de clasificación de falsos positivos. En este aspecto, cobra relevancia contar con textos correctamente normalizados y lograr capturar, en la recolección de información, palabras específicas que permitan caracterizar una actividad económica, a fin de distinguir entre clases afines. Por ejemplo, en CIUO-1 se observó un desempeño relativamente bajo en las clases 1 y 3, que puede ser explicado por el hecho de que estas clases contienen actividades que se refieren a un perfil técnico o profesional del informante que están siendo capturados por la clase 2.

Al igual que en muchas otras aplicaciones de *machine learning*, declarar un algoritmo como el mejor para la clasificación de glosas es una tarea difícil y, quizás, casi imposible. Los experimentos y análisis realizados en este estudio, sin embargo, han revelado algunas características interesantes entre los tres métodos investigados. Se resumen a continuación:

1. Naïve Bayes (NB): es simple y el más rápido en el aprendizaje de modelo entre los tres métodos. Puede funcionar bien para la clasificación de textos. Dado que el algoritmo asume que las características individuales son completamente independientes entre sí, el modelo puede beneficiarse de una selección de características efectiva, lo que se ha demostrado en los experimentos realizados. En la misma línea, NB puede presentar un bajo rendimiento si se aplica sobre un conjunto de datos donde hay algunas dependencias observables entre las características. En los experimentos efectuados se observó que NB tuvo un rendimiento sobresaliente en términos de *precision* sobre la clase más poblada en cada uno de los clasificadores. Una posible explicación es el hecho de haber establecido las probabilidades *a priori* a través de las frecuencias observadas en los documentos.

2. Support Vector Machine (SVM): como se ha reportado en muchos estudios previos, SVM es un clasificador muy estable y también es escalable a la dimensionalidad de características. En este trabajo, SVM se desempeña como el mejor clasificador, reflejado a través de las medidas de $recall$, $precision$, F_1 y $recall/precision$. El punto de equilibrio $recall/precision$ de SVM fue de 86,6% para CIUO-1, 85,3% para CIUO-2, 95,3% para CAENES-1 y 93,0% para CAENES-2. La función *kernel* y el parámetro de costo *C* seleccionados para SVM influyen dramáticamente en su resultado. En este trabajo se optó por una función *kernel lineal,* que resultó ser relativamente rápida (aproximadamente 2 horas en procesamiento) en el entrenamiento de modelos, mientras que el parámetro de costo fue determinado a través de *cross – validation,* con 10 *folds*. SVM presenta una importante propiedad respecto a la forma en que llega a la mejor función de clasificación, estableciendo un margen de separación máximo entre dos clases. Esto le confiere a SVM una poderosa capacidad de generalización en la clasificación.

3. Random Forest (RF): es el método que mejor funciona en la clasificación de CIUO-2 y es muy competitivo, teniendo desempeños similares a los obtenidos con SVM. Si bien el entrenamiento de los modelos es relativamente lento (aproximadamente 14 horas en procesamiento) debido a la exigencia de *hardware* en sus cálculos, la implementación *bagging* que caracteriza a RF le confiere una poderosa capacidad de generalización en la clasificación. Dado que el número de características a utilizar en RF es crucial, en este trabajo se optó por un número recomendado en la literatura y configurado por *default* en la implementación R, es decir, raíz cuadrada del número de características. Este parámetro puede ser calibrado en un futuro para obtener mejores resultados.

Los tres métodos evaluados funcionan muy bien en la tarea de clasificación de documentos, destacando el desempeño observado para CAENES, donde alcanzan un 90% en todas las métricas consideradas. Una posible explicación de este notable resultado es el hecho de que la glosa usada para el entrenamiento (y prueba) de los modelos se desprende solamente de la descripción del "sector económico" de la empresa donde trabaja el informante, lo que proporciona textos más cortos, con palabras específicas que caracterizan una actividad económica. Por otra parte, los métodos tienen tiempos de procesamiento razonables, sin

embargo, pueden ser disminuidos mediante la mejora en los recursos de *hardware* y *software* disponibles.

Es claro que los tres algoritmos son alternativas viables para obtener clasificaciones precisas y en tiempos mucho menores a los reportados por la clasificación manual, lo que operacionalmente reduce los costos y aumenta la eficiencia en la labor de la institución.

SVM presenta en términos estadísticos los mejores desempeños, aunque RF es una buena alternativa para la clasificación de CIUO-2 por tener una capacidad similar al de SVM. Sin embargo, en términos computacionales, SVM ofrece un menor tiempo de procesamiento. En consecuencia, a través de los criterios estadísticos y computacionales analizados, se recomienda el uso del modelo de SVM para la clasificación de las glosas de la Encuesta Nacional de Empleo.

6. Conclusiones y proyecciones

Un típico proceso de clasificación de textos consiste en los siguientes pasos: preprocesamiento, reducción de dimensionalidad, representación adecuada de documentos y clasificación. En este trabajo han sido descritos diferentes métodos para todos estos pasos, presentando alternativas que pueden ser probadas con mayor profundidad en nuevos estudios. Además de la introducción teórica de estos métodos, se mostró su aplicación para la clasificación de las glosas de la ENE, evaluando los métodos NB, SVM y RF. Estos algoritmos fueron evaluados utilizando el conjunto de datos 2017 de la ENE, obteniendo mejores resultados con el método SVM, el que además es de fácil aplicación. En el futuro, los métodos utilizados pueden ser mejorados a través de una calibración más fina de los parámetros de los modelos, además de la mejora en la calidad en los datos de entrada, lo que podría generar aún mejores resultados en términos de precisión.

Este documento responde a la necesidad de contar con sistemas automatizados para la clasificación de los grandes volúmenes de información que maneja el INE, y que actualmente se clasifican, en su mayoría, de forma manual. Así, los algoritmos de *text mining* y *machine learning* introducidos en este estudio se proponen como una alternativa viable para obtener clasificaciones más precisas y mucho menos costosas para la institución.

Con el dramático incremento del uso de internet y otras fuentes de información, ha ocurrido una explosión del volumen de documentos de interés que las instituciones requieren administrar. En esta tarea, el INE debe tomar un rol fundamental, adoptando adecuadamente las últimas tecnologías disponibles para el desarrollo de su rol público. Así, en el futuro la institución debe estudiar la aplicación de estos algoritmos u otros de mayor complejidad (como los algoritmos de *deep learning*) en volúmenes de datos mucho mayores.

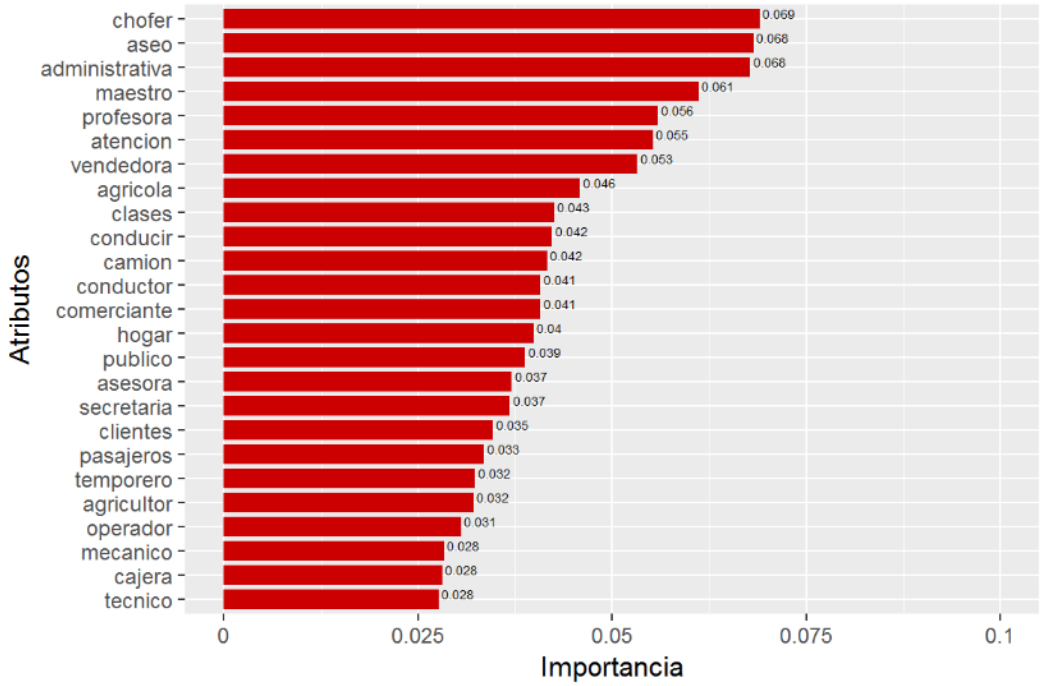
Finalmente, este trabajo demuestra que es posible, en el contexto de la labor pública, generar mejoras sobre la base de la utilización de *software* libre y de código abierto, como la plataforma R, lo que podría liberar recursos que hoy en día se utilizan en productos de pago, además de aprovechar las posibilidades de colaboración entre instituciones que este tipo de *software* facilitan.

Anexos

Anexo 1: Ganancia de Información

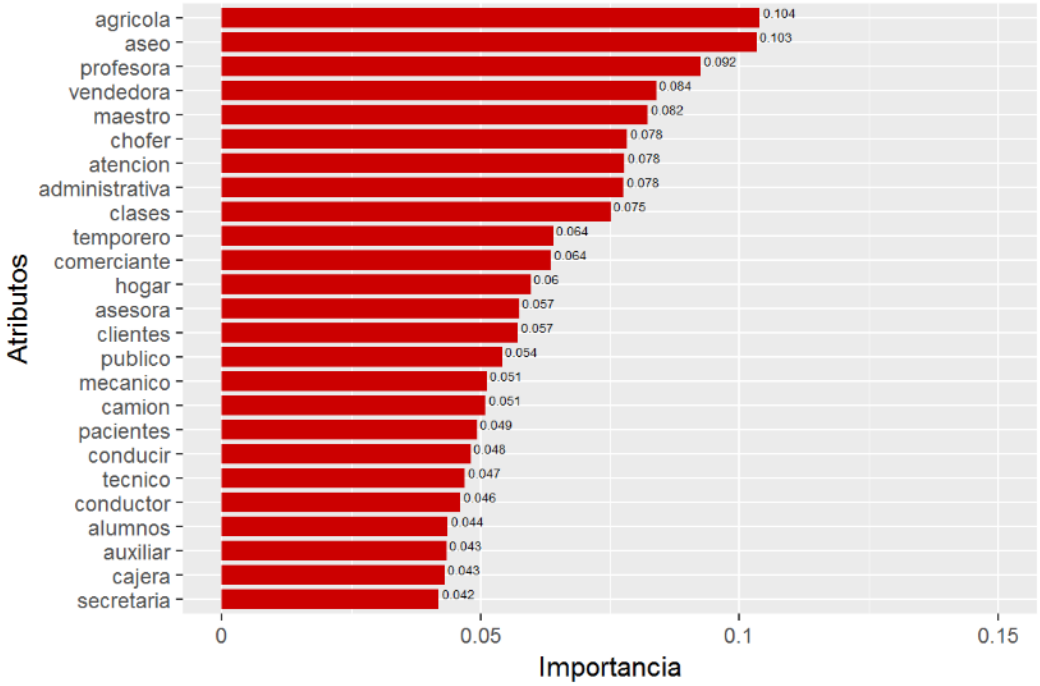
En los gráficos A.1. a A.4. se muestra la Ganancia de Información (IG) del *top* 25 de atributos (características) para los clasificadores CIUO-1, CIUO-2, CAENES-1 y CAENES-2, respectivamente.

Gráfico A.1. Ganancia de Información de *top* 25 de atributos para CIUO – 1



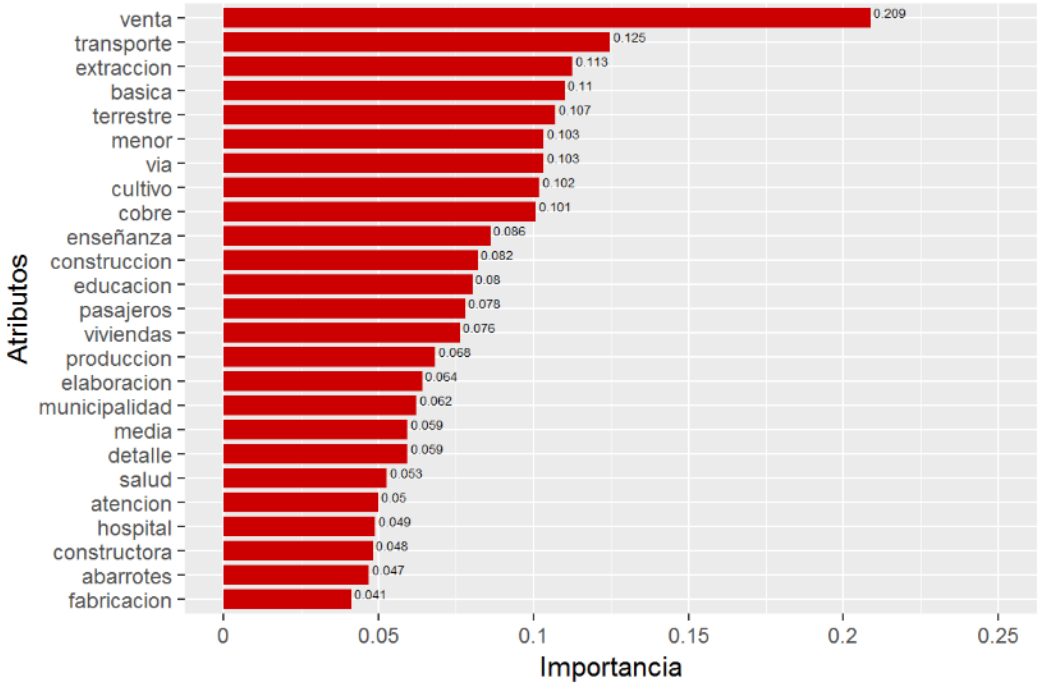
Fuente: Elaboración propia.

Gráfico A.2. Ganancia de Información de *top* 25 de atributos para CIUO – 2



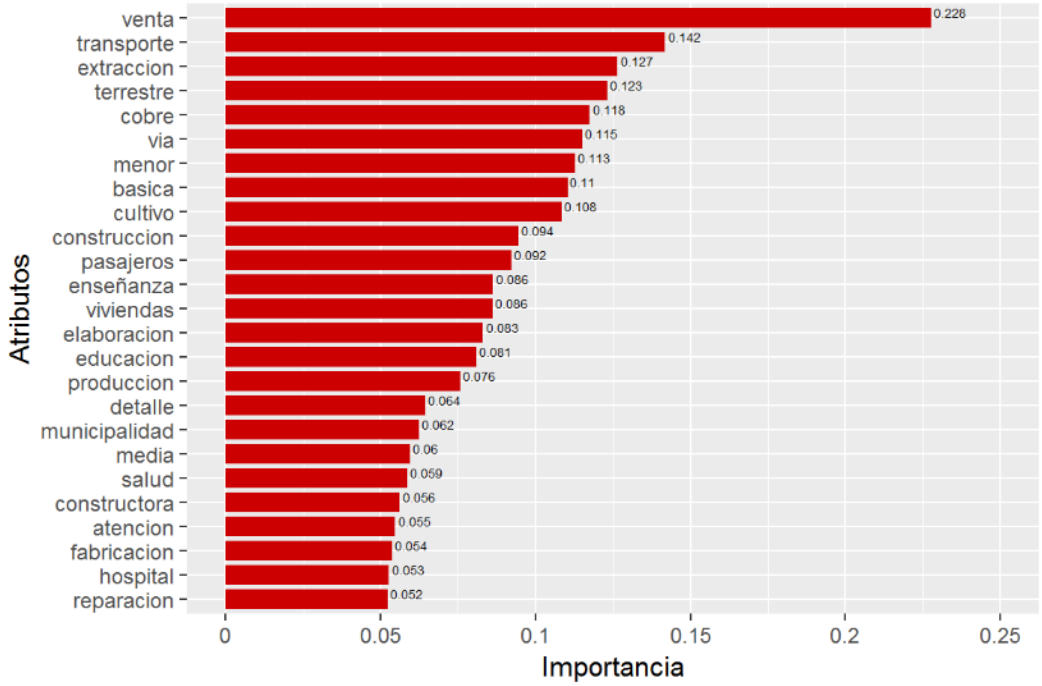
Fuente: Elaboración propia.

Gráfico A.3. Ganancia de Información de *top* 25 de atributos para CAENES – 1



Fuente: Elaboración propia.

Gráfico A.4. Ganancia de Información de *top* 25 de atributos para CAENES – 2



Fuente: Elaboración propia.

Anexo 2: Resultados de CAENES – 2

En la tabla A.1. se muestran en orden descendente los números de documentos en la colección usados para el entrenamiento y prueba de los modelos para cada una de las clases que componen el CAENES-2.

Tabla A.1. Conjuntos de datos de entrenamiento y prueba para CAENES – 2

Clase	No. entrenamiento	No. prueba
48. Comercio al por mayor y al por menor, excepto de vehículos automotores y motocicletas	23.908	5.848
1. Agricultura, ganadería, caza y actividades de servicios conexas	14.653	3.684
85. Enseñanza	13.027	3.201
84. Administración pública y de defensa; planes de seguridad social de afiliación obligatoria	9.064	2.343
41. Construcción de edificios	6.775	1.727
86. Actividades de atención de la salud humana	6.657	1.715
49. Transporte por vía terrestre y transporte por tuberías	6.652	1.703
56. Actividades de servicio de comidas y bebidas	5.221	1.280
10. Elaboración de productos alimenticios	5.071	1.202
43. Actividades especializadas de construcción	3.796	874
4. Extracción y procesamiento de cobre	3.743	977
45. Comercio y reparación de vehículos automotores y motocicletas	2.810	708
96. Otras actividades de servicios personales	2.077	468
52. Almacenamiento y actividades de apoyo al transporte	1.740	466
81. Actividades de servicios a edificios y de paisajismo	1.663	411
14. Fabricación de productos textiles, fabricación de prendas de vestir, productos de cuero y productos conexos	1.624	426
69. Actividades jurídicas y de contabilidad	1.610	384
3. Pesca, acuicultura y actividades de servicios conexas	1.528	399
64. Actividades de servicios financieros, excepto las de seguros y fondos de pensiones	1.486	373
2. Silvicultura, extracción de madera y actividades de servicios conexas	1.388	356

Clase	No. entrenamiento	No. prueba
25. Fabricación de productos elaborados de metal y servicios de trabajo de metales, excepto máquinas y equipos	1.301	316
16. Producción de madera y fabricación de productos de madera y corcho, excepto muebles; fabricación de artículos de paja y de materiales trenzados	1.256	342
94. Otras actividades de servicios personales	1.216	288
55. Actividades de servicio de comidas y bebidas	1.177	295
68. Actividades inmobiliarias	1.127	288
42. Comercio y reparación de vehículos automotores y motocicletas	1.014	239
95. Reparación de computadores y de efectos personales y enseres domésticos	965	241
61. Telecomunicaciones	935	249
93. Actividades deportivas, de esparcimiento y recreativas	830	222
88. Actividades de asistencia social sin alojamiento	817	198
71. Actividades de arquitectura e ingeniería; ensayos y análisis técnicos	816	211
35. Suministro de electricidad, gas, vapor y aire acondicionado	806	197
97. Actividades de los hogares como empleadores de personal doméstico	769	187
33. Mantenimiento, reparación e instalación de maquinaria y equipo	679	167
11. Elaboración de bebidas alcohólicas y no alcohólicas	642	180
36. Captación, tratamiento y distribución de agua	618	131
31. Fabricación de muebles	587	179
87. Actividades de asistencia social en instituciones	556	135
82. Actividades administrativas y de apoyo de oficina y otras actividades de apoyo a las empresas	531	176
62. Actividades de programación y consultorías informáticas y otras actividades conexas	519	136
65. Seguros, reaseguros y fondos de pensiones, excepto planes de seguridad social de afiliación obligatoria	513	128
23. Fabricación de otros productos minerales no metálicos	503	103
17. Fabricación de papel y productos de papel	494	129

Clase	No. entrenamiento	No. prueba
22. Fabricación de productos de caucho y de plástico	425	114
20. Fabricación de sustancias y productos químicos	419	97
80. Actividades de seguridad e investigación	410	114
74. Otras actividades profesionales, científicas y técnicas	400	97
32. Otras industrias manufactureras	385	113
90. Actividades creativas, artísticas y de entretenimiento	383	100
8. Explotación de otras minas y canteras	367	88
18. Impresión y reproducción de grabaciones	361	93
73. Publicidad y estudios de mercado	349	74
92. Actividades de juegos de azar y apuestas	324	71
38. Recogida, tratamiento y eliminación de desechos; recuperación de materiales	312	85
50. Transporte por vía acuática	302	101
66. Actividades auxiliares a los servicios financieros y a los seguros	295	74
53. Actividades postales y de mensajería	283	78
7. Extracción de minerales metalíferos, excepto cobre	282	65
77. Actividades de alquiler y arrendamiento, excepto inmuebles	273	63
24. Fabricación de metales comunes	268	65
72. Investigación científica y desarrollo	214	52
21. Fabricación de productos farmacéuticos, sustancias químicas medicinales y productos botánicos de uso farmacéutico	203	34
51. Transporte por vía aérea	202	40
30. Fabricación de vehículos automotores, remolques y semirremolques y otros tipos de equipo de transporte	189	58
60. Actividades de programación y difusión de radio y televisión	184	41
78. Actividades relacionadas con el suministro de empleo	182	44

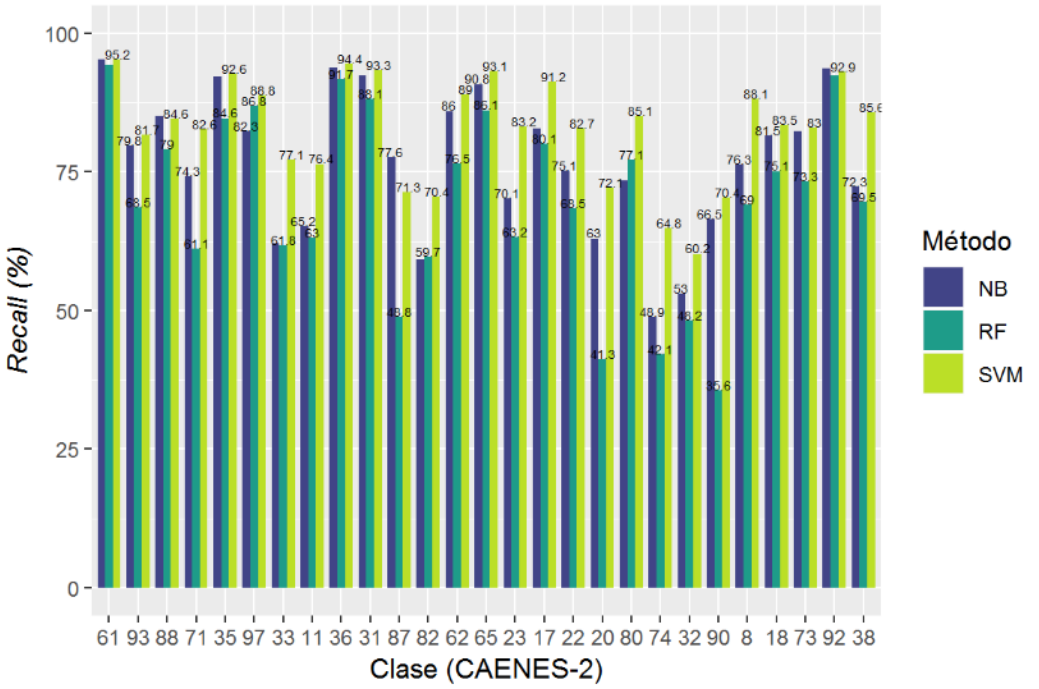
Clase	No. entrenamiento	No. prueba
27. Fabricación de productos informáticos, electrónicos, ópticos, equipos eléctricos y maquinaria	181	43
79. Actividades de agencias de viajes, operadores turísticos, servicios de reservas y actividades conexas	177	55
9. Actividades de servicio de apoyo para la explotación de minas y canteras	169	29
59. Actividades de producción de películas cinematográficas, videos y programas de televisión, grabación de sonido y edición de música	154	44
58. Actividades de edición	150	36
75. Actividades veterinarias	142	29
70. Actividades de oficinas principales, actividades de consultoría de gestión	134	38
91. Actividades de bibliotecas, archivos y museos y otras actividades culturales	110	27
6. Extracción de petróleo crudo y gas natural	96	22
19. Fabricación de coque y productos de la refinación del petróleo	77	21
5. Extracción de carbón de piedra y lignito	50	12
37. Evacuación de aguas residuales	43	14
99. Actividades de organizaciones y órganos extraterritoriales	24	5
12. Elaboración de productos de tabaco	23	6
63. Actividades de servicios de información	18	6

Fuente: Elaboración propia.

En los gráficos A.5. a A.10. se muestran los desempeños de los tres métodos en términos de *recall*, *precision* y medida F_1 para las clases de CAENES-2 ubicadas entre la posición 28 a la 81[3] según el número de documentos de entrenamiento (y de prueba) en la colección.

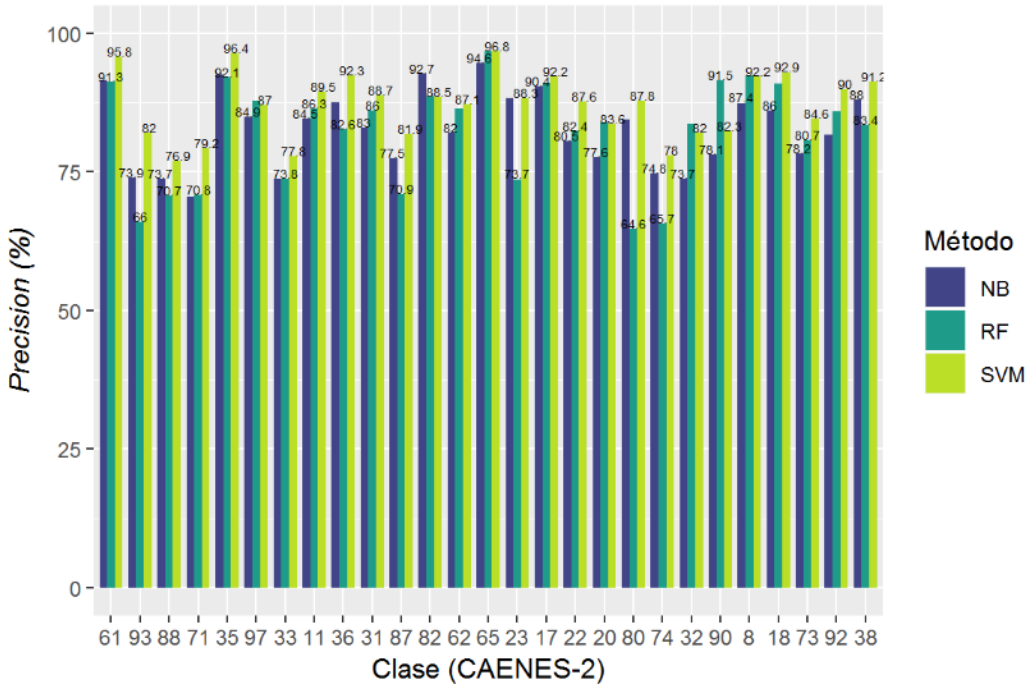
[3] El clasificador CAENES-2 contiene 83 clases, sin embargo, a partir del conjunto de datos usado solo fue posible reportar indicadores para 81 de ellas.

Gráfico A.5. Desempeño de *Recall* según método para CAENES – 2. Clases en posiciones 28 a 54 según número de documentos en la colección



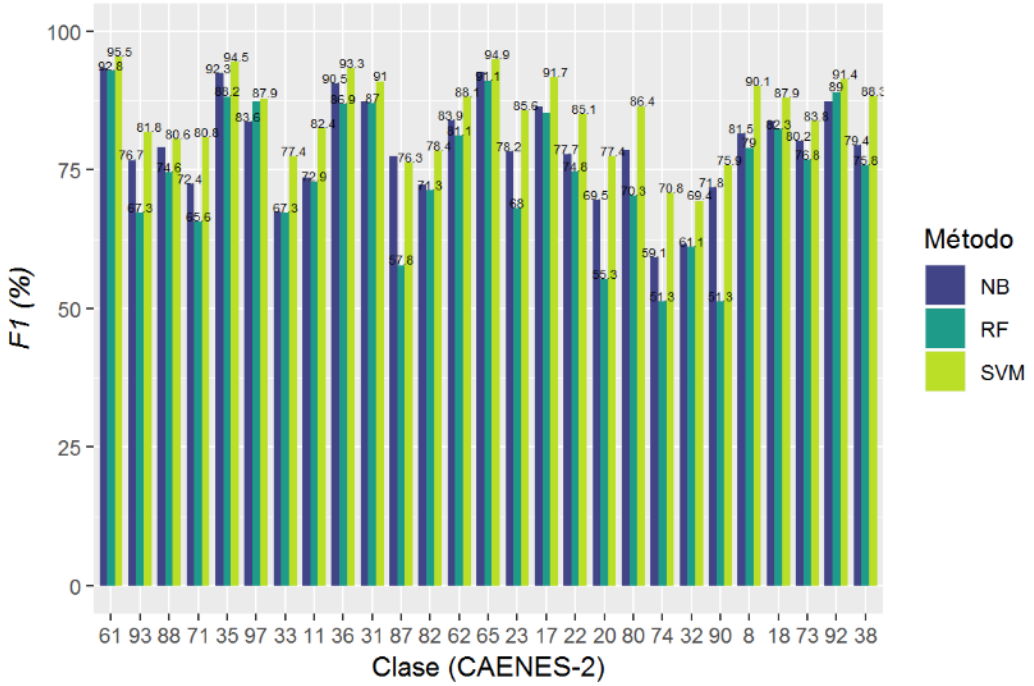
Fuente: Elaboración propia.

Gráfico A.6. Desempeño de *Precision* según método para CAENES – 2. Clases en posiciones 28 a 54 según número de documentos en la colección



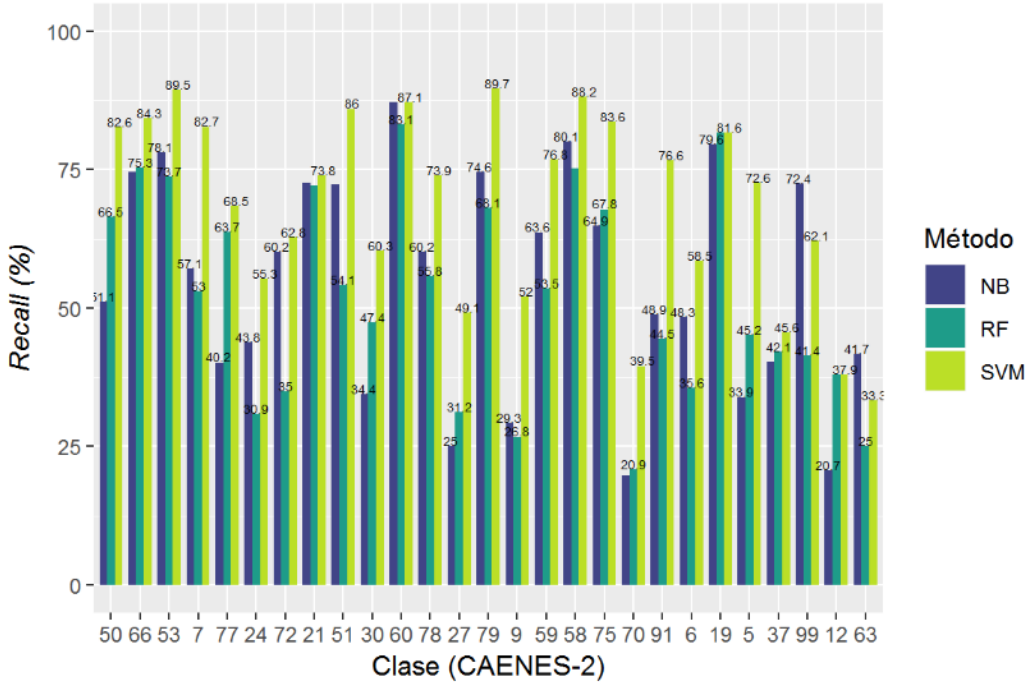
Fuente: Elaboración propia.

Gráfico A.7. Desempeño de la medida F_1 según método para CAENES – 2. Clases en posiciones 28 a 54 según número de documentos en la colección



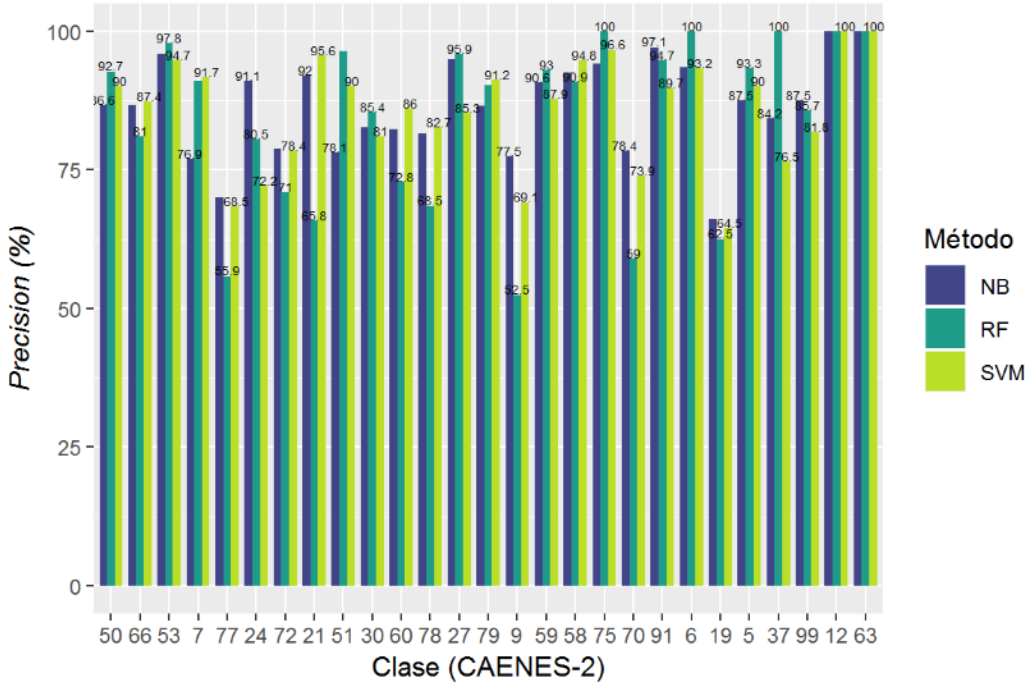
Fuente: Elaboración propia.

Gráfico A.8. Desempeño *Recall* según método para CAENES – 2. Clases en posiciones 55 a 81 según número de documentos en la colección.



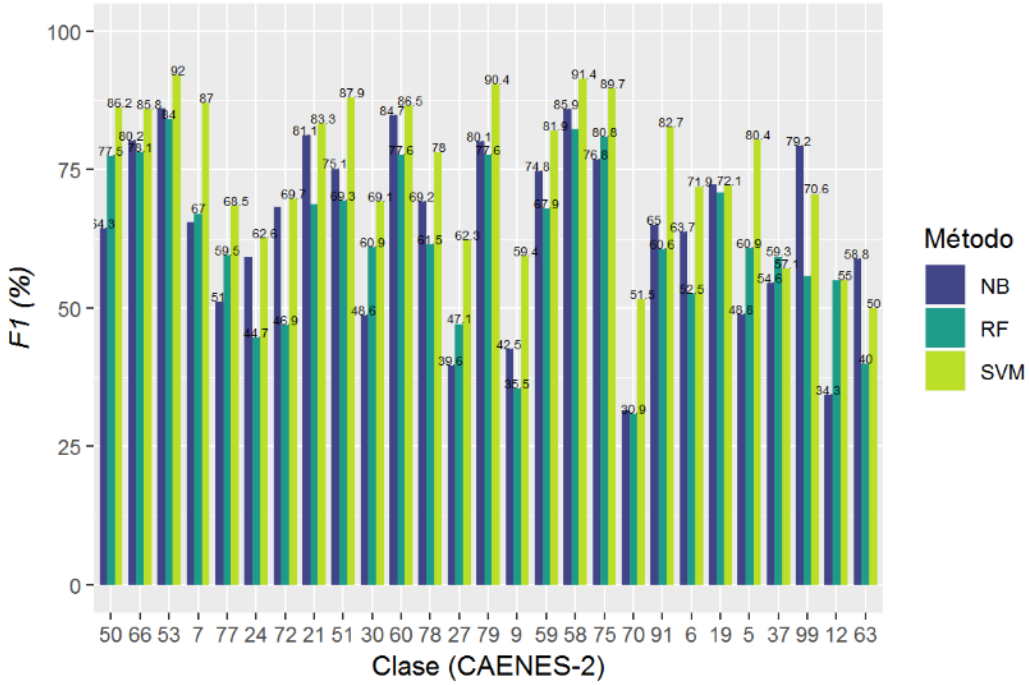
Fuente: Elaboración propia.

Gráfico A.9. Desempeño *Precision* según método para CAENES – 2. Clases en posiciones 55 a 81 según número de documentos en la colección.



Fuente: Elaboración propia.

Gráfico A.10. Desempeño de la medida F_1 según método para CAENES – 2. Clases en posiciones 55 a 81 según número de documentos en la colección.



Fuente: Elaboración propia.

Bibliografía

Aas, K. & Eikvil, L. (1999). *Text Categorisation: A Survey*, Norwegian Computer Center, 3 – 37.

Alfaro, R. & Allende, H. (2011). *Text Representation in Multi-label Classification: Two New Input Representations*, International Conference on Adaptive and Natural Computing Algorithms 2011, 1-10.

Berry, M. W. & Kogan, J. (2010). Text Mining: Applications and Theory. United Kingdom: John Wiley & Sons, Ltd.

Breiman, L. (2001). *Random Forests*, Machine Learning, 45 (1), 5 – 32.

Departamento Administrativo Nacional de Estadística (DANE), Colombia (Diciembre de 2005). *Clasificación Internacional Uniforme de Ocupaciones Adaptada para Colombia*. Obtenido desde https://www.dane.gov.co/files/sen/nomenclatura/ciuo/CIUO_88A_C_2006.pdf

Instituto Nacional de Estadísticas (INE), Chile (Abril de 2016). *Clasificador de Actividades Económicas Nacional para Encuestas Sociodemográficas (CAENES)*. Obtenido desde http://historico.ine.cl/canales/chile_estadistico/mercado_del_trabajo/empleo/metodologia/pdf/caenes.pdf

Joachims, T. (1998). *Text Categorization with support vector machines: Learning with many relevant features*, In Proc. 10th European Conference on Machine Learning (ECML), Springer Verlag.

Liu, Ch., Chan, Y., Alam, S.H. & Fu, H. (2015). *Financial Fraud Detection Model: Based on Random Forest*, International Journal of Economics and Finance 7 (7), 178 – 188.

Manning, CH., Raghavan, P. & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Mitchell, T.M. (1997). Machine Learning. McGraw-Hill Science.

Ooms, J. (2017). *High Performance Stemmer, Tokenizer, and Spell Checker*. Obtenido desde https://cran.r-project.org/web/packages/hunspell/hunspell.pdf

Pérez, S. (2017). *Análisis y Clasificación de Textos con Técnicas Semi Supervisadas Aplicado a Área Atención al Cliente* (tesis de pregrado). Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sebastiani, F. (2002). *Machine Learning in Automated Text Categorization*, ACM Computing Surveys, 34(1), 1 – 47 .

Welbers, K., Van Atteveldt, W. & Benoit, K. (2017). *Text Analysis in R*, Communication Methods and Measures, 11(4), 245 – 265 .

Yang, Y. & Pedersen, J. (1997). *A Comparative Study on Feature Selection in Text Categorization*, Proceedings of the Fourteenth International Conference on Machine Learning, 412 – 420.