SISTEMA DE CLASIFICACIÓN Y CODIFICACIÓN AUTOMÁTICA EN LA ENCUESTA NACIONAL DE EMPLEO

DEPARTAMENTO DE ESTUDIOS LABORALES

SUBDIRECCIÓN TÉCNICA

INSTITUTO NACIONAL DE ESTADÍSTICAS

Mayo/2019

ÍNDICE

RESUMEN

En este documento se describen los elementos del Sistema de Clasificación[1] y Codificación[2] Automática desarrollado en el INE para abordar las necesidades de codificación de preguntas abiertas contenidas dentro del proceso de producción estadística. En particular, este documento se enfoca en la Encuesta Nacional de Empleo (ENE) del INE, cuyo propósito es la elaboración de estadísticas oficiales sobre ocupación y desocupación, la cual enfrenta la necesidad puntual de codificar cinco preguntas de su cuestionario que describen con textos las respuestas de los informantes en cuanto a la actividad económica de la empresa/establecimiento y la ocupación de la población objetivo. A pesar del foco en la ENE, la metodología que aquí se describe es extensible para la elaboración de estadísticas cuya fuente de información corresponda a textos abiertos con descripciones de las mismas temáticas, es decir, ocupación y/o actividad económica. Para ejecutar la labor de clasificar dichos textos en la categoría correspondiente, este trabajo se basa en las definiciones y criterios de los clasificadores internacionales; Clasificador Internacional Industrial Uniforme (CIIU) y Clasificador Internacional Uniforme de Ocupaciones (CIUO), combinado con técnicas de minería de textos y de aprendizaje de máquinas que, en conjunto con la data codificada históricamente en un proceso manual y el aprendizaje adquirido de éste, ha permitido desarrollar la metodología de este sistema de clasificación automática.

[1] Para este trabajo se entiende como el acto de identificar un texto con una categoría determinada.
[2] Para este trabajo se entiende como el acto de asignar el código correspondiente a un texto clasificado.

INTRODUCCIÓN

Este trabajo describe los aspectos centrales del Sistema de Clasificación y Codificación Automática desarrollado para abordar la necesidad de clasificación de los textos que se pesquisan en las Encuestas del Instituto Nacional de Estadísticas (INE). En particular, para este trabajo se consideró la clasificación de los textos derivados de la captura en la Encuesta Nacional de Empleo (ENE) de las variables "oficio, labor u ocupación" y del "sector económico donde trabaja" dentro de la población ocupada y cesante, cuyas respuestas son declaradas por los informantes, junto con el resto de las preguntas del cuestionario de la ENE. El proceso histórico de clasificación contemplaba el análisis de los textos por parte de un equipo de codificadores en gabinete que determinaban el código correcto según dicta el estándar internacional correspondiente.

Los textos históricos codificados de manera manual se utilizaron como insumo base para desarrollar un sistema que, a través de minería de textos y aprendizaje computacional, predice matemáticamente la clasificación correspondiente a cada texto. Los distintos niveles de complejidad para la clasificación se originan, en el caso de la ocupación, debido a que no basta con la glosa de la ocupación para su clasificación, sino, según estándar OIT, se necesita utilizar información de las tareas y las competencias para desempeñar la ocupación. En el caso de la clasificación de actividad económica, muchas veces, se presentan textos que describen actividades que representan a más de un establecimiento y, por el contrario, también se presentan textos con información insuficiente para poder clasificar la actividad económica. Para abordar esto último, históricamente se ha podido contar con la información complementaria que brinda la ENE en sus módulos de caracterización laboral, lo que, por cierto, también forma parte del proceso de clasificación automático que aquí se describe.

Este trabajo propone un cambio en el enfoque tradicional de lo que implica la codificación manual propiamente tal, a una que utiliza la metodología automática basada en un modelo óptimo definido por el INE para el acto de clasificar el texto, y que se focaliza en analizar los resultados derivados de la aplicación de este proceso sistemático, de una evaluación y depuración continua del proceso para fines de publicación, así como la necesaria actualización del modelo incorporando los nuevos hallazgos en términos de nuevas actividades económicas u ocupaciones que surjan en el mercado laboral. Esto último debido a que, cualquier metodología automática de clasificación de textos está sujeta a la sensibilidad de las palabras que componen la descripción de un cierto código de clasificación y, por tanto, esto impacta en los resultados del modelo. También, debido a la información complementaria que posee la ENE y la implementación de los procesos de auditoría permanente en la clasificación en la encuesta, se genera un espacio de introducción de mejora continua a la data que se publica y que al mismo tiempo alimenta la máquina de aprendizaje de los modelos de codificación automática. El análisis se realiza una vez ejecutados los algoritmos de clasificación y, se lleva a cabo por parte de expertos en clasificadores, lo cual permite ofrecer un proceso basado en un sistema de aprendizaje supervisado con resultados más precisos, que no presenta los problemas manuales de codificación en múltiples códigos, lo que mejora

considerablemente los resultados al reducir los errores no muestrales propios del proceso de codificar[3].

El desarrollo de este trabajo se fundamenta en tres grandes pilares. El primero, los clasificadores internacionales de sector económico y ocupación denominados CIIU y CIUO, respectivamente, los que brindan todas las categorías a identificar en cada caso que corresponda, además de las directrices y criterios para la clasificación propiamente tal. En particular, se utilizan las adaptaciones nacionales de los clasificadores, a saber, CAENES y CIUO 08.CL en niveles de 1 y 2 dígitos. Este último publicado oficialmente por el INE a fines de 2018. El segundo pilar corresponde a la metodología de clasificación automática de textos de Support Vector Machine (SVM) que, basado en un aprendizaje computacional de los casos codificados manualmente durante 2017 y 2018 en la ENE, permitió automatizar el proceso de codificación de los textos que describen las actividades y ocupaciones declaradas por los informantes en la encuesta. El tercer pilar corresponde al control de calidad que retroalimenta al sistema, a través de la actualización de casos de entrenamiento para los modelos derivados del SVM, los que permiten mejorar los resultados como consecuencia del análisis de casos críticos e incorporar las dinámicas propias del mercado, lo que puede verse reflejado, por ejemplo, en la introducción de plataformas tecnológicas como Rappi, Uber, Glovo, etc., que implica la aparición de nuevas palabras en los textos pesquisados en el trabajo de campo si comparamos con lo observado hace algunos años atrás.

A nivel institucional, el sistema de Codificación Automática se encuentra alineado en el marco de los ejes estratégicos del INE, ya que permite instalar estándares de calidad en la producción estadística nacional con aplicación en productos del Sistema Estadístico Nacional (SEN), pues se mejora la eficacia de la codificación, entregando mayor precisión en la codificación, así como mitigando sesgos en el uso de los clasificadores de forma manual, además de reducir significativamente los tiempos de producción, incluso en grandes volúmenes de textos[4]. La ejecución de modelos SVM ha sido desarrollado como un servicio compartido de disposición libre para las encuestas que lo requieran, lo cual lo enmarca en el eje de excelencia organizacional, ya que además de la eficacia, mejora la eficiencia en el uso de recursos incorporando la modernización a través del uso de herramientas tecnológicas y de acceso libre en el INE.

CAPITULO I: PROBLEMÁTICA

La producción estadística de la ENE y otras fuentes de información, bajo el esquema business statistical generalized process model (BSGPM), enfrentan una fase en su proceso productivo denominado “procesamiento”, el que contempla la labor de “clasificar y codificar” como un subproceso principal. Para el caso específico de la ENE, se enfrenta la necesidad de leer los textos declarados por los informantes que describen la ocupación y la actividad económica de la población

[3] Véase sección 6 de resultados, página 25.

[4] En la ENE los tiempos de codificación manual contemplaban más de 3.500 horas de trabajo al mes, distribuidas en un equipo de 22 personas durante todo el mes. El método automático requiere menos de 4 horas de trabajo para su ejecución.

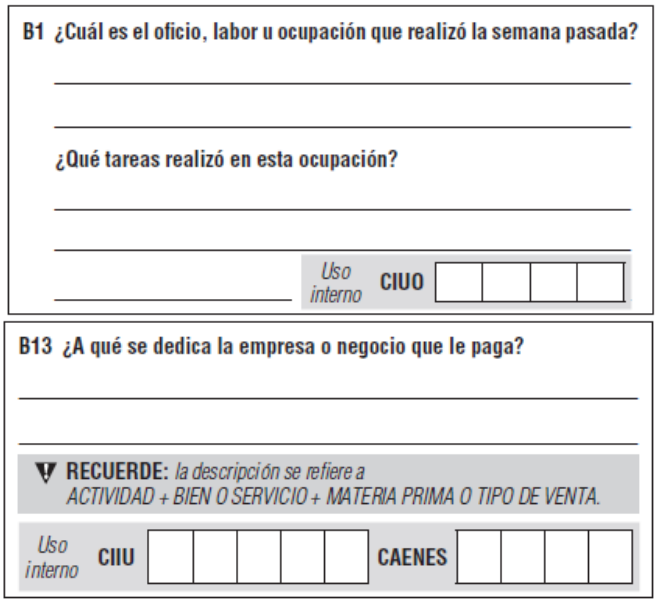
ocupada y clasificarlos según los estándares internacionales que el INE adopta en materias de clasificación en determinados grupos de ocupación o determinados sectores de la economía.

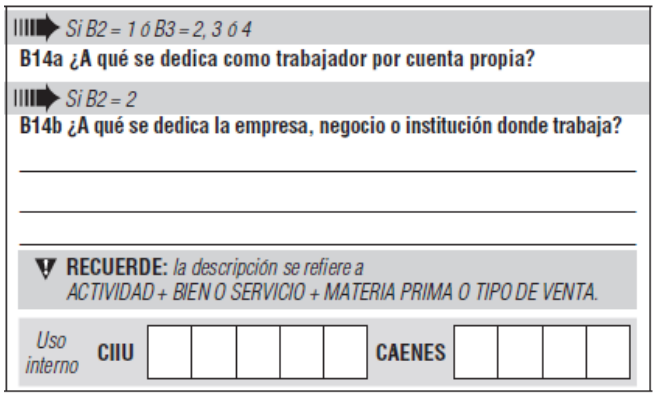
Para la clasificación de un texto en un determinado grupo ocupacional (1° dígito), así como en un determinado subgrupo principal (2° dígito), se utiliza el estándar CIUO 08.CL, sin perjuicio de que también fue probado para la anterior clasificación CIUO 88, solo para fines de investigación interna, ya que este clasificador se deja de publicar a contar del trimestre febrero-abril 2019, siendo reemplazado por la versión CIUO 08.cl. Los textos que requieren ser clasificados se denominan glosas y representan lo descrito por los informantes en las preguntas del cuestionario de la ENE, que pesquisan el oficio, labor u ocupación realizada durante un período de referencia, junto con las tareas desempeñadas en la misma.

Para la clasificación de actividad económica, se dispone del clasificador CAENES, publicado por el INE en abril 2016, que corresponde a la adaptación nacional para encuestas sociodemográficas (o de hogares) del Clasificador Internacional Industrial Uniforme, el que brinda las secciones (1° dígito) y divisiones (2° dígito) que abarcan todos los sectores de la economía donde deben ser clasificadas las actividades declaradas por los informantes. Para este propósito, la ENE pesquisa, específicamente, la actividad que desarrolla una persona ocupada como cuenta propia o bien la actividad del establecimiento, negocio o institución donde trabaja, en caso de ser un trabajador dependiente. Para los casos de ocupados subcontratados o vinculados a través de suministro, se pesquisa también la información de la empresa que le paga.

Ambas temáticas pesquisadas, se preguntan a la población ocupada (módulo B del cuestionario) y a la población cesante (módulo E del cuestionario). La especificación de las preguntas se puede ver a continuación a través de las imágenes del cuestionario de la ENE, así como sus respectivas respuestas más frecuentemente observadas, que en definitiva representan el insumo al cual se ejecuta el acto de clasificar y codificar.

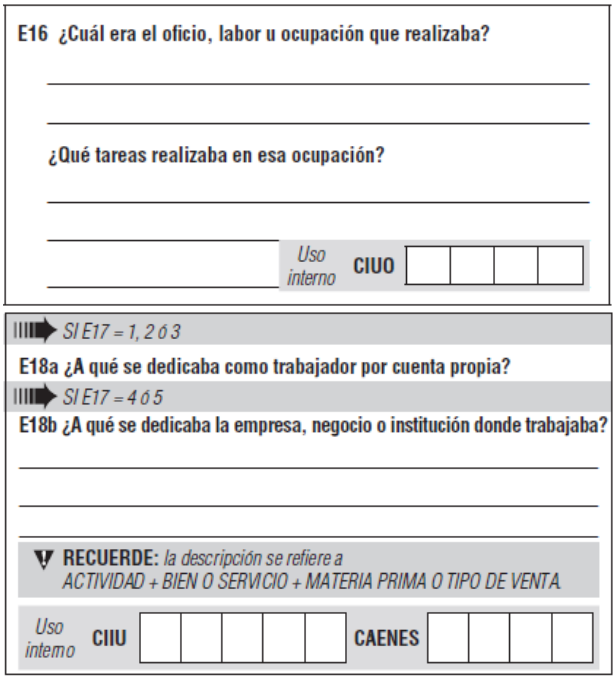
Set 1: Preguntas a Ocupados que requieren codificación en base a clasificadores internacionales





Fuente: Cuestionario ENE

Set 2: Preguntas a Cesantes con codificación en base a clasificadores internacionales



Fuente: Cuestionario ENE

Tabla 1: Textos/Glosas más frecuentes de la pregunta sobre actividad económica. Hombres, trimestre diciembre-febrero 2019[5].

Ranking	Glosa	Frecuencia	Ranking	Glosa	Frecuencia
1	extraccion cobre	334	**16**	transporte pasajeros via terrestre urbano	43
2	construccion viviendas	207	**17**	cultivo arandanos	40
3	municipalidad	159	**18**	mineria extraccion cobre	40
4	carabineros chile	149	**19**	constructora viviendas	39
5	extraccion mineral cobre	111	**20**	empresa constructora viviendas	39
6	ejercito chile	95	**21**	cria engorda ganado bovino	37
7	produccion leche cruda vaca	77	**22**	transporte carga pesada	37
8	armada chile	73	**23**	cultivo uva campo abierto	36
9	transporte pasajeros via terrestre	63	**24**	aserradero madera	35
10	extraccion refinacion ventas cobre	60	**25**	condominio habitacional	35
11	mineral extraccion cobre	54	**26**	gendarmeria chile	31
12	supermercado minorista	49	**27**	puerto comercial	31
13	transporte pasajeros	49	**28**	construccion viviendas particulares	30
14	banco privado	48	**29**	hospital	30
15	ministerio obras publicas	48	**30**	ilustre municipalidad	29

Fuente: ENE.

Tabla 2: Textos/Glosas más frecuentes de la pregunta sobre actividad económica. Mujeres, trimestre diciembre-febrero 2019.

Ranking	Glosa	Frecuencia	Ranking	Glosa	Frecuencia
1	municipalidad	132	**16**	sala cuna jardin infantil	29
2	jardin infantil	100	**17**	educacion prebasica basica	28
3	hospital	89	**18**	establecimiento educacion basica	28
4	supermercado minorista	67	**19**	restaurant	28
5	cultivo arandanos	55	**20**	centro salud familiar	26
6	hospital publico	52	**21**	clinica salud privada	26
7	banco comercial	44	**22**	extraccion cobre	26
8	banco privado	44	**23**	construccion viviendas	25
9	educacion basica	40	**24**	educacion basica prebasica	25
10	jardin infantil sala cuna	39	**25**	enseñanza basica basica	25
11	ventas abarrotes	35	**26**	universidad privada	25
12	enseñanza prebasica basica	34	**27**	enseñanza basica	23
13	ventas abarrotes menor	33	**28**	supermercado	23
14	carabineros chile	31	**29**	callcenter	22
15	supermercado comercio minorista	31	**30**	cultivo arandanos aire libre	22

Fuente: ENE.

[5] En todas las tablas que se presentan textos/glosas más frecuentes, éstas omiten la ortografía del castellano, así como palabras irrelevantes para la clasificación, tales como artículos definidos, indefinidos, etc. Lo anterior debido al tratamiento propio de los textos con fines de separación de las palabras contenidas en las glosas.

Tabla 3: Textos/Glosas más frecuentes de la pregunta sobre ocupación. Hombres, trimestre diciembre-febrero 2019.

Ranking	Glosa	Frecuencia	Ranking	Glosa	Frecuencia
1	temporero agricola	495	**16**	auxiliar aseo	156
2	obrero agricola	489	**17**	vendedor	156
3	guardia seguridad	483	**18**	soldador arco	139
4	maestro carpintero	391	**19**	labores agricolas	131
5	agricultor	365	**20**	trabajador agricola	123
6	administrativo	286	**21**	carpintero	120
7	comerciante	284	**22**	jornal	115
8	chofer camion	273	**23**	operario produccion	115
9	jardinero	236	**24**	comerciante ambulante	110
10	mecanico automotriz	219	**25**	conductor camion	110
11	bodeguero	182	**26**	pescador artesanal	103
12	maestro albañil	176	**27**	conserje	99
13	temporero fruticola	175	**28**	supervisor	98
14	maestro construccion	171	**29**	garzon	97
15	comerciante establecido	165	**30**	mecanico	93

Fuente: ENE.

Tabla 4: Textos/Glosas más frecuentes de la pregunta sobre ocupación. Mujeres, trimestre diciembre-febrero 2019.

Ranking	Glosa	Frecuencia	Ranking	Glosa	Frecuencia
1	asesora hogar	1269	**16**	educadora parvulos	125
2	auxiliar aseo	652	**17**	vendedora meson	123
3	comerciante	503	**18**	comerciante ambulante	122
4	administrativa	443	**19**	peluquera	122
5	manipuladora alimentos	428	**20**	aseadora	121
6	vendedora	345	**21**	comerciante establecida	116
7	secretaria administrativo	334	**22**	tecnico paramedico	103
8	cajera	303	**23**	enfermera	94
9	temporera agricola	300	**24**	asistente social	84
10	secretaria	231	**25**	cocinera	83
11	tecnico enfermeria	184	**26**	tecnico parvulos	79
12	temporera fruticola	163	**27**	reponedora	78
13	costurera	161	**28**	comerciante establecido	77
14	garzona	158	**29**	guardia seguridad	76
15	ayudante cocinar	135	**30**	niñera	76

Fuente: ENE.

CAPITULO II: MINERÍA DE TEXTO

La inmensa cantidad de información textual que circula en Internet ha provocado el desarrollo de métodos, algoritmos y sistemas capaces de realizar el procesamiento y análisis de textos estructurados, semi estructurados y no estructurados con fines de organización, clasificación y análisis de estos. Así, surge la minería de texto como un área de estudio, o bien como disciplina interdisciplinaria según la literatura que la define como una, en donde participan semiólogos, matemáticos, lingüistas, sociólogos, entre otros (Contreras, 2014).

En términos generales, la minería de textos es el proceso de descubrimiento de patrones y nuevos conocimientos a partir de un cúmulo de textos no necesariamente relacionados previamente. También es definida como un proceso para analizar textos o enunciados para extraer información que resulta útil para propósitos particulares (Contreras, 2014)

La minería de texto es el descubrimiento no trivial potencialmente útil de conocimiento partiendo de una colección de documentos no estructurado, análogamente al descubrimiento[6].

Así la minería de texto y todo su desarrollo en técnicas de clasificación automática es una de las áreas de investigación que ha cobrado mayor importancia en los últimos años, debido a que conjuga el análisis de los grandes volúmenes de textos digitales que se almacenan en bases de datos empresariales, páginas web y redes sociales, con la necesidad de generar información relevante a partir de los mismos.

La minería de texto ha permitido el desarrollo y aplicación de máquinas de aprendizaje para clasificación de textos, una línea de la inteligencia artificial y computacional que se basa en el desarrollo de algoritmos que “aprenden” o reconocen patrones recurrentes en cada clase observada, a partir de grandes volúmenes de textos de entrada, previamente clasificados por humanos. A su vez, un programa computacional aprende y su desempeño será mejor según la experiencia previa.

A pesar de los avances en la sofisticación de la clasificación automática de información heterogénea con gran precisión, la clasificación automática de textos por lo general, sigue siendo un proceso supervisado, lo que se traduce en que se requiere que especialistas clasifiquen y ordenen la información para poder nutrir y entrenar al sistema, para que en definitiva con esta información, el clasificador automático sea capaz de generar una clasificación inédita y replicarlo con distinta información (Cárdenas, Olivares y Alfaro, 2014).

En términos de la fuente de origen del insumo, la ENE es una encuesta que actualmente se levanta en hogares en cuestionario de papel, por lo que los registros con los textos pasan por un proceso de digitación manual, en el cual no existe homogeneidad en el tratamiento de abreviaciones, ortografía y tratamiento en general de los textos (caracteres, puntuaciones, etc.), tal que deriva en una necesidad de normalizar los textos, o lo que también se denomina proceso de cleandata. En definitiva, todos los textos pasan por un proceso de eliminación de caracteres especiales (no alfabéticos), eliminación de signos de puntuación, eliminación de números y eliminación de espacios

[6] Sánchez, D., Martín-Bautista, M. “Un enfoque deductivo para la minería de texto”, 2006.

en blanco múltiple. También se aplica una fase de homologación de palabras que se escriben diferente, pero representan una misma ocupación o actividad, tales como homologaciones de género, por ejemplo. Posteriormente, la técnica requiere de un proceso de tokenización, que consiste en separar la totalidad del texto de entrada en características más específicas conocidas como tokens, las que corresponden específicamente a las palabras o combinaciones de éstas que constituyen los componentes semánticos más significativos de los textos.

Una aplicación práctica de esta técnica para fines del Sistema de Clasificación y Codificación Automática, es la identificación de palabras individuales (tokenización) que se observan en las glosas y como éstas se vinculan a las categorías del clasificador respectivo según lo codificado. Con esta acción, que se ejecuta a partir de la librería "tm", más "wordcloud", permite generar una de las imágenes más características en el campo de esta disciplina, correspondiente a las nubes de palabras.

Nube de palabras 1: Palabras de las glosas de actividad económica declaradas por hombres.

Fuente: Respuestas tokenizadas de la pregunta sobre rama de actividad económica en la ENE, trimestre diciembre-febrero 2019.

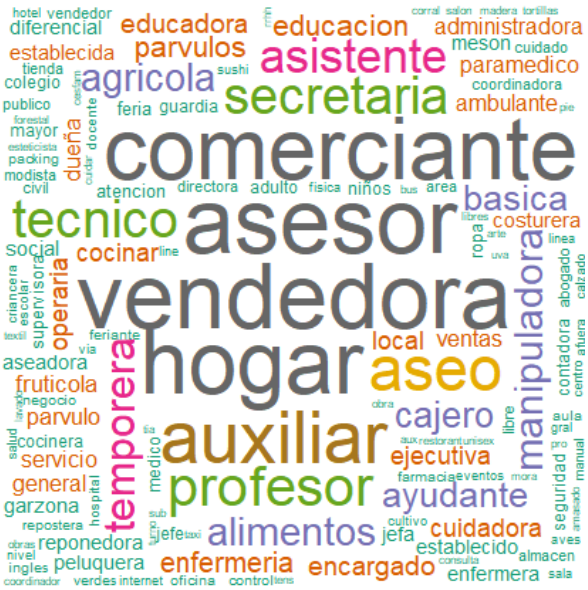
Nube de palabras 2: Palabras de las glosas de actividad económica declaradas por mujeres.

Fuente: Respuestas tokenizadas de la pregunta sobre rama de actividad económica en la ENE, trimestre diciembre-febrero 2019.

Nube de palabras 3: palabras de las glosas de ocupación declaradas por hombres.

Fuente: Respuestas tokenizadas de la pregunta sobre ocupación en la ENE, trimestre diciembre-febrero 2019.

Nube de palabras 4: Palabras de las glosas de ocupación declaradas por mujeres.



Fuente: Respuestas tokenizadas de la pregunta sobre ocupación en la ENE, trimestre diciembre-febrero 2019.

Al realizar una primera interpretación de los datos recientemente expuestos, es posible visualizar cómo se reflejan las evidentes diferencias que se presentan en las ocupaciones entre hombres y mujeres. Sin embargo, en el caso del sector económico, destaca que, tanto en hombres como en mujeres, la palabra más repetida en sus glosas es “ventas”. Dada la sensibilidad de cualquier método de clasificación automática a situaciones como esta, se hace imperioso auditar los casos a través de minería de texto para la exploración de subgrupos de glosas, como por ejemplo las codificadas en cada clase, permitiendo de este modo poder desarrollar un proceso integral de clasificación de textos descriptivos de la ocupación de una persona y del sector económico donde trabaja.

De esta técnica también deriva el concepto de matriz documento-término, la cual corresponde a la representación vectorial de los textos/glosas y se utiliza para tratamiento matemático de la data. En términos prácticos, corresponde a que los textos que describen ocupación o actividad económica son representados por vectores de palabras. En consecuencia, una matriz de un set de N casos de la ENE, estará compuesta por el total de glosas únicas en las filas y, el total de palabras de la data en las columnas. Lo importante de esto es que permite cuantificar los pesos relativos de las palabras individuales en cada glosa, así como a través de la totalidad de las glosas, pues estas medidas son fundamentales para poder realizar predicciones sobre distintas combinaciones de palabras, o sea, predecir la clasificación de un texto/glosa cualquiera, a partir de la información observada.

CAPITULO III: CLASIFICACIÓN AUTOMÁTICA CON MÁQUINAS DE SOPORTE VECTORIAL (SUPPORT VECTOR MACHINES [SVM])7

3.1. Determinación modelo óptimo de clasificación automática

Durante el otoño de 2018, en un trabajo conjunto entre profesionales de los departamentos de Estudios Laborales e Investigación y Desarrollo del INE, se generó la instancia de utilizar los textos de ocupación y sector económico pesquisados en la ENE durante abril de 2015 y diciembre de 2017[8], con el objetivo de utilizar esta data como base para el entrenamiento de modelos de aprendizaje de máquinas, tal que se cumpliera a su vez un segundo propósito de mejorar las técnicas de clasificación automática desarrolladas previamente en el INE[9]. De este modo, se dispuso de la información proporcionada por un total de 505.958 registros de personas ocupadas en el período de estudio, a partir de la cual los profesionales de Investigación y Desarrollo del INE Julio Guerrero y Julián Cabezas desarrollaron la metodología que determinó un modelo óptimo para la clasificación automática de los textos de la ENE con técnicas de minería de textos. Específicamente, el trabajo en comento utilizó los casos disponibles en la ENE durante el período de estudio, lo que significó que se basó en los clasificadores CIUO 88 y CAENES.

Para la determinación de los modelos de clasificación automática de grupo principal (1° dígito) y subgrupo principal (2° dígito), con fundamento en CIUO, se utilizó el texto especificado en la pregunta B1 del cuestionario de la ENE. Esta, pesquisa el oficio, labor u ocupación realizada durante la semana de referencia, junto con las tareas desempeñadas en la misma.

Para la clasificación de actividad económica con CAENES a nivel de sección (1° dígito) y división (2° dígito), se utilizaron los textos de la pregunta B14 del cuestionario, la cual refleja la descripción de la persona ocupada en cuanto a la actividad que desarrolla como cuenta propia o la actividad del establecimiento, negocio o institución donde trabaja.

Se probaron las técnicas de Naïve Bayes (NB), Random Forest (RF) y Support Vector Machines (SVM), las cuales consideran las características y la relación entre los textos, las palabras que lo componen y sus respectivos códigos aplicados en el proceso manual histórico. De este modo, el propósito es replicar las estimaciones de la población objetivo (original), según ocupación y actividad económica. En definitiva, el ejercicio consiste en aprender de la codificación manual aplicada históricamente, para a partir del 80% de los casos, entrenar un modelo que permita predecir con mejor resultado los códigos de los textos del 20% restante de casos que son utilizados como base de prueba. En

[7] En esta sección se describe una breve síntesis de los aspectos más centrales de la técnica, pues los detalles se pueden revisar directamente en el documento "Clasificación automática de textos utilizando técnicas de text mining: Aplicación a las glosas de la Encuesta Nacional de Empleo (ENE), J. Guerreo & J. Cabezas, INE. Publicado en https://www.ine.cl/docs/default-source/documentos-de-trabajo/metodologicos/clasificacion-automatica-de-textos-utilizando-tecnicas-de-text-mining-aplicacion-a-las-glosas-de-la-encuesta-nacional-de-empleo.pdf?sfvrsn=0

[8] El período se determinó en base a la información disponible a la fecha, con clasificador CAENES.

[9] En particular, para mejorar la experiencia de clasificación automática desarrollada para el Censo 2017 que utilizó métodos econométricos y matemáticos.

consecuencia, el propósito es emular de mejor manera los códigos asignados de manera manual, pues esta medida del grado de coincidencia entre los casos codificados manual y automáticamente es utilizada como indicador del rendimiento de las técnicas en estudio. Su cuantificación se realiza para los casos de ocupados en cada categoría del clasificador respectivo.

Ahora bien, como se señaló en la sección anterior, la medida de rendimiento estará más o menos sesgada dependiendo de la calidad de la información utilizada para el entrenamiento. Para el caso de la ENE, es importante señalar que, debido a la heterogeneidad en la aplicación de criterios de manera manual, la base de entrenamiento presenta casos con múltiples códigos asignados a un mismo texto, -se presenta con mayor frecuencia en la clasificación de ocupaciones y con escasa frecuencia en sector económico-, lo que limitaría las conclusiones del rendimiento efectivo del modelo. Para subsanar esta situación, una vez determinado el modelo óptimo, se procedió a auditar casos con fines de homologación de criterios y limpieza de bases de entrenamiento.

Con un grado de rendimiento entre 85% y 95%, dependiendo del clasificador y el nivel (1 o 2 dígitos), SVM resultó ser la técnica óptima para la clasificación de textos que describen la ocupación y el sector económico donde se desempeña la población ocupada.

El SVM, es un tipo de algoritmo usado en la clasificación automática de textos, y corresponde a máquinas de aprendizaje que toman distintas características de los elementos que se quieren clasificar y los llevan a un espacio vectorial multidimensional. Es en este espacio, donde el algoritmo identifica, de forma óptima, un hiperplano que separa a los vectores de una clase del resto. Es en ese concepto de 'separación óptima' donde reside la característica fundamental de estos algoritmos, asociado a una serie de propiedades teóricas y prácticas atractivas (Vapnik, 1995 en Cárdenas, Olivares y Alfaro, 2014)".

La ventaja de los modelos SVM, es que logran identificar una frontera decisión lineal entre dos clases, a través de una línea que los separe, maximizando el espacio en el hiperplano, a partir de una función llamada Kernel, la cual permite realizar separaciones no lineales de los datos, proyectando la información a un espacio de características de mayor dimensión (Figura 1). En términos sencillos, los modelos SVM superan la distinción netamente binaria, agregando características al vector (instrucciones que se transforman en "aprendizaje") en un hiperplano, permitiendo generar nuevas clasificaciones y generando relaciones difíciles de detectar en grandes volúmenes de información. Es relevante señalar, que el SVM en la literatura es considerado un método de clasificación automática semi asistido o supervisado, ya que requiere de un "entrenamiento", vale decir, de la indicación de reglas y códigos de base, para poder generar nuevas formas de clasificación y, en la medida que la base de entrenamiento refleje mejor estas reglas, mejor será el resultado.

En cuanto a clasificación de textos, se observa un amplío desarrollo en técnicas de clasificación estadística y machine learning. Así por ejemplo destaca el uso de modelos de regresión, clasificadores basados en vecinos más cercanos, árboles de decisión, clasificadores Bayesianos, algoritmos de aprendizaje de reglas, Support Vector Machines y redes neuronales, entre otras.

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

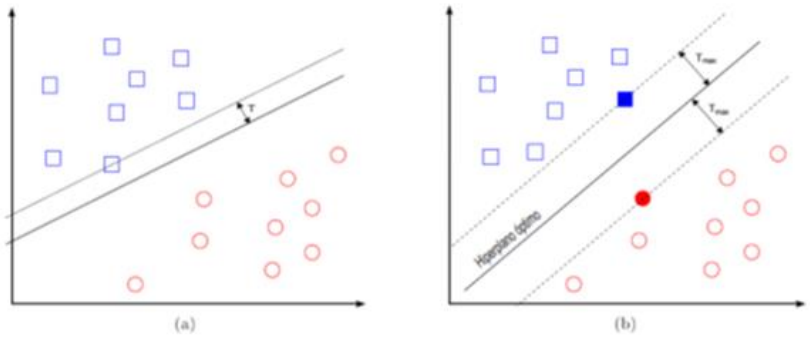
dimensión y el aprendizaje de la tarea de clasificación en ese espacio sin ninguna complejidad computacional adicional se logran mediante el uso de una función de Kernel.

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 mientras 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.

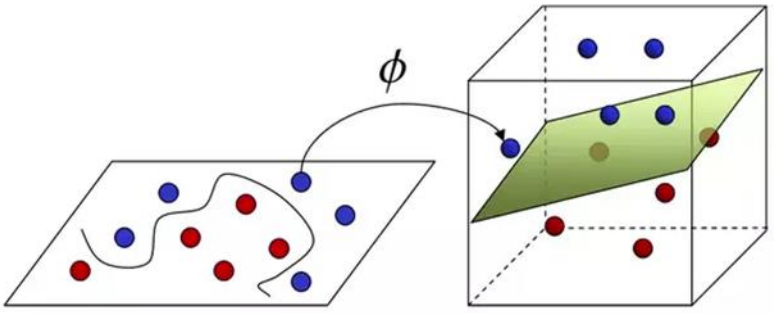
Esto es aplicable solo para tareas de clasificación binaria, por lo que para usar este método para la clasificación de textos (problema de clases múltiples) tiene que ser tratado como una serie de problemas de clasificación dicotómica.

Figura 1: Separación lineal de elementos por clasificar



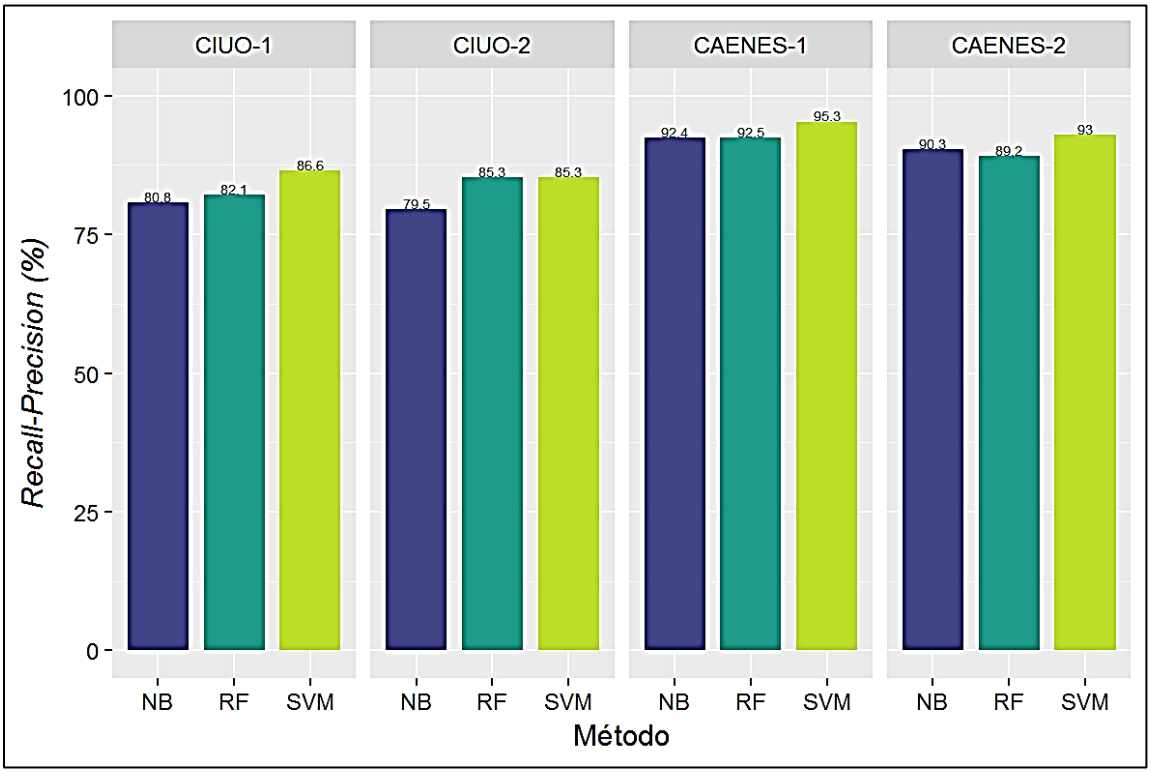
Fuente: Nashif, Shadman & Rakib Raihan, Md & Rasedul Islam, Md & Imam, Mohammad. (2018).

Figura 2: Separación multidimensional de elementos por clasificar



Fuente: Nashif, Shadman & Rakib Raihan, Md & Rasedul Islam, Md & Imam, Mohammad. (2018).

Figura 3: Desempeño de Recall-Precision según método para Clasificadores CIUO 88 y CAENES



Fuente: "Clasificación automática de textos utilizando técnicas de text mining", J. Guerrero, INE, 2019.

En términos prácticos, el problema consiste en encontrar los vectores que maximizan la distancia de los márgenes que separan las características entre dos categorías diferentes, en un contexto de múltiples características. Entonces, a través de la función de Kernel, que permite tratar el problema de manera lineal, se deduce el parámetro ϕ que permite identificar de forma óptima un hiperplano, el que separa a los vectores de una clase con el resto. En consecuencia, siguiendo el caso de CAENES, la técnica determina que la clasificación de un texto se evalúa como un problema de selección entre dos categorías, hasta completar la evaluación, de a dos en dos, entre los 21 sectores económicos especificados en el clasificador.

Del proceso deriva que cada palabra pesquisada en las glosas es ponderada según su frecuencia en cada categoría, en este caso sería el sector económico, y también según su frecuencia entre las distintas glosas observadas en el mismo sector. A partir de ambas características, se determina la probabilidad de que una glosa pertenezca a una categoría, en función de las mismas palabras que componen la glosa que se requiere clasificar.

3.2. Modelos SVM requeridos para la ENE

Las necesidades concretas de codificación de la ENE están dadas por las 5 preguntas abiertas en su cuestionario referidas a ocupación y actividad económica de la respectiva población objetivo. Como se señaló en la sección 2, hay dos preguntas de ocupación (B1 y E16) que deben ser codificadas con

base en CIUO 08.CL para generar datos de ocupación bajo el nuevo clasificador de ocupaciones adaptado a la realidad nacional y que reemplaza al CIUO 88.

La clasificación en la ENE se requiere a nivel de gran grupo (GG), pero para efectos de aplicación de la técnica determinada, se trabajó también a nivel de subgrupo principal (SP), entendiendo ciertamente eso sí, que esta medida no corresponde a una estimación oficial de la encuesta sino más bien un insumo intermedio para lograr mayor precisión en la clasificación a 1 dígito.

También se necesita abordar las tres preguntas de actividad económica (B13, B14 y E18), las cuales deben ser codificadas con CAENES.

Los modelos SVM utilizados para codificar aplicando el clasificador CAENES corresponde a la adopción directa de los modelos óptimos desarrollados por Guerrero y Cabezas 2019, pero dado que estos fueron determinados con datos de entrenamiento del año 2017 y se requiere aplicar el método oficialmente a los casos pesquisados en la ENE a partir de la submuestra de abril de 2019, se actualizó la data de entrenamiento con las glosas de enero a diciembre de 2018.

Para la clasificación de ocupaciones con clasificador CIUO 08.CL, se adoptó el mismo método óptimo antes señalado, pero se desarrolló a partir de una base de entrenamiento proveniente de otra fuente de información, puesto que la ENE no disponía de casos codificados manualmente con el clasificador en comento, pues hasta el momento de este estudio, solo contaba con la codificación del clasificador antiguo. Dado esto, se tuvo que utilizar información proveniente de la VIII Encuesta de Presupuestos Familiares (EPF), además de un número importante de glosas codificadas y auditadas especialmente para fines de construir el modelo.

Es importante señalar que los modelos SVM señalados cuentan con una plataforma web, en donde estos son alojados y, en conjunto con los elementos derivados del entrenamiento que permiten la acción de clasificar los textos, ésta se ejecuta de manera autónoma, sin necesidad de requerir software R Studio, a pesar de que los algoritmos estén desarrollados en esta plataforma. Dicha plataforma web fue desarrollada por el INE y su propósito es que, a partir de los datos de la ENE, el proceso de codificar puede ser requerido a este sistema compartido que en definitiva ejecuta un script R con las instrucciones del proceso que solicita las glosas, las procesa y las retorna con su respectivo código. En la siguiente sección se especifican más detalles sobre esto.

CAPITULO IV: SISTEMA DE CLASIFICACIÓN Y CODIFICACIÓN AUTOMÁTICA

Considerando que la ENE es una encuesta de carácter continuo, el sistema comienza por determinar el entrenamiento de la máquina, con datos de un periodo previo (en este caso todo 2018), el que se deberá aplicar a los datos de un periodo siguiente (por ejemplo, un trimestre móvil de 2019). Cada vez que se ejecuta esta actividad, derivan los insumos que se cargan al servicio compartido para la ejecución de los modelos SVM determinados sobre los nuevos textos digitados durante el período en curso. Los insumos son: Matriz de documento-término, Modelo SVM y el script que instruye las fases del proceso de clasificación propiamente tal.

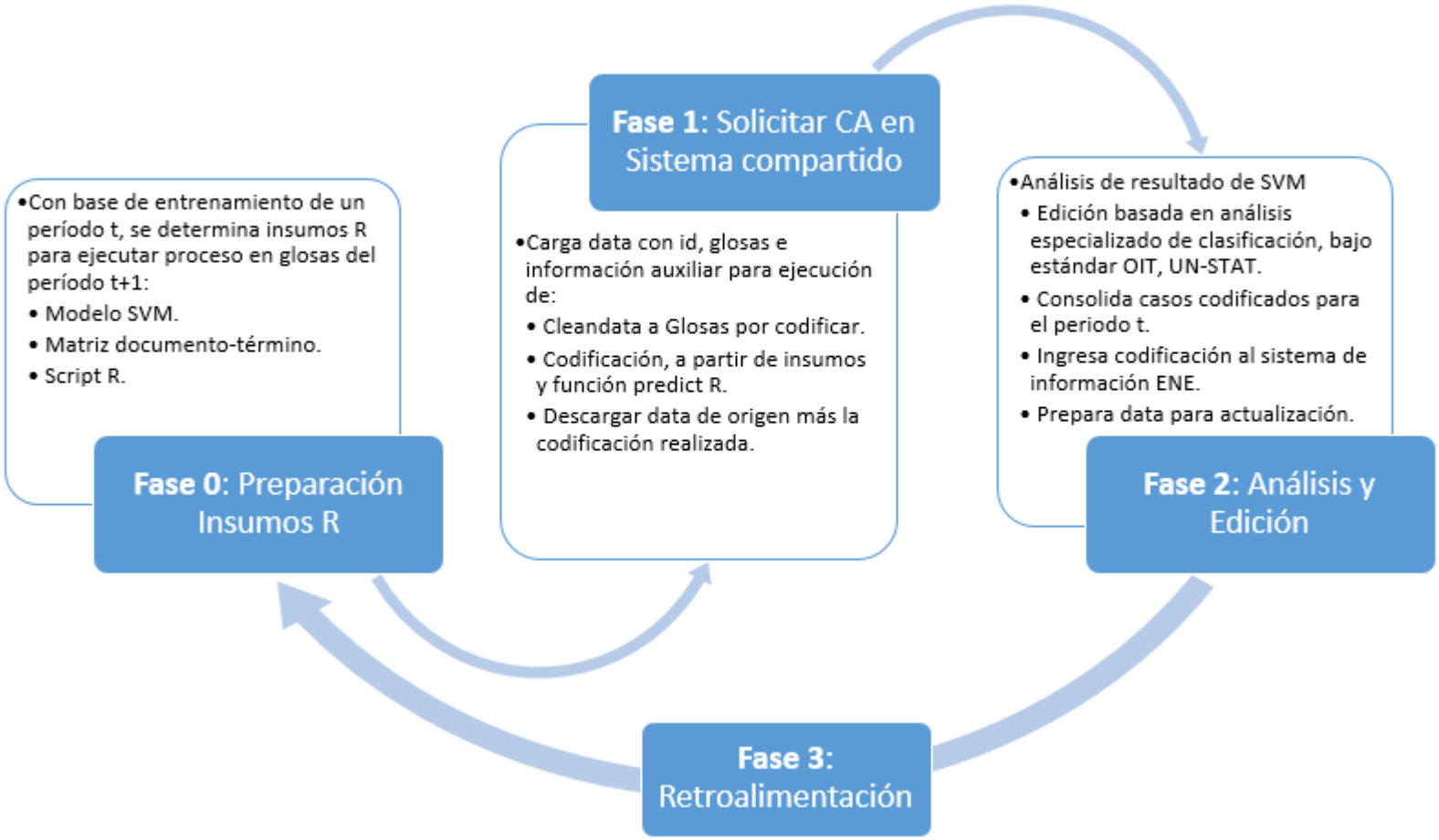
El proceso continúa con el acto de solicitar el servicio de codificación en el portal web desarrollado para este fin (Servicio compartido). Este requerirá las glosas a codificar y un identificador del caso en formato JSON[10], verificará la data, transforma la data en formato de tabla para en definitiva ejecutar el proceso instruido en el script R, el cual aplicará una limpieza de los textos con fines de eliminación de caracteres irrelevantes y de homologación semántica, para finalmente ejecutar el modelo SVM que clasifica los casos requeridos. Las glosas ingresadas al sistema serán devueltas con su respectivo código, más la probabilidad con que fueron clasificadas en dicha categoría.

El proceso continúa con la fase de análisis ex-post que consiste en supervisar permanente y focalizadamente los casos codificados con SVM, con fines de control de calidad para la publicación mensual de las estimaciones derivadas con la encuesta, así como para la retroalimentación del modelo, a través de la constante actualización de las bases de entrenamiento requerida para la determinación del modelo SVM a utilizar en el próximo período (Ver diagrama 1).

Para efectos de producción en la ENE, el sistema contempla la metodología óptima descrita anteriormente, pero ajustando siempre la data de entrenamiento a los últimos casos disponibles a la fecha, por tanto, la base de entrenamiento original se actualizará en los periodos siguientes:

[10] Formato de texto. Del inglés JavaScript Object Notation.

Figura 4: Proceso operativo de Clasificación y Codificación Automática ENE



4.1. Descripción detallada del proceso

Fase 0: Entrenamiento y Estimación Modelo SVM

La fase 0 contempla la preparación de los insumos centrales para ejecutar el acto de clasificar y codificar los textos. Siguiendo a Guerrero y Cabezas 2019, se actualizó la data a 2018 para el modelo estimado para clasificar con CAENES, al tiempo que para clasificar CIUO, se desarrolló un modelo que sigue la misma metodología pero que utiliza información clasificada según la nueva versión del clasificador, a saber, el CIUO 08.CL.

Del proceso de estimación del modelo, derivan también la matriz documento-término[11] y el modelo SVM estimado. Ambos son los insumos que se requieren para aplicar una función de predicción que permite clasificar y codificar los casos requeridos.

Debido a la necesidad permanente de actualización, esta fase contempla una actualización periódica que deriva del análisis especializado en clasificadores.

Tabla 5: Características de los modelos para actividad y ocupación

Data Entrenamiento	Variable Entrenada	Clasificador	Nivel	Costo óptimo	Número de glosas	Rendimiento
ENE Año 2018	Actividad económica	CAENES	Sección	0,75	177.774	95,86%
ENE Año 2018	Actividad económica	CAENES	División	0,75	177.774	94,52%
EPF VIII + ENE	Ocupación	CIUO08.CL	Grupo	4	27.688	94,93%
EPF VIII + ENE	Ocupación	CIUO08.CL	Subgrupo	5,25	27.688	98,75%

Fuente: Elaboración propia.

Fase 1: Utiliza Sistema Compartido de Codificación Automática

A nivel institucional, el desarrollo de este trabajo fue posicionado estratégicamente, con lo que se accedió al desarrollo de un servicio web que permite la ejecución del proceso de codificación automática a través de un proceso centralizado y autónomo.

Para su ejecución se requiere cargar en el sistema los insumos derivados del entrenamiento de los modelos, más las respectivas glosas a codificar. De este modo, el servicio web ejecuta un script R en donde se realiza el proceso de:

i. Cleandata de nuevas glosas por codificar.
ii. Tokenización de glosas por codificar.

[11] Específicamente su dimensionalidad.

iii. Ejecución de función de predicción a partir de; data requerida, matriz documento-término y modelo SVM.

Fase 2: Análisis especializado para edición de SVM

Este ítem contempla el análisis especializado de clasificadores aplicado a los resultados de la codificación automática con SVM.

La fortaleza de este ítem radica en el análisis con uso de información auxiliar proporcionada por la misma encuesta, o bien cuando, por ejemplo, existen glosas que "por denominación" van en cierto código[12]. Esto último toma relevancia sobre todo en los casos en que se transita de un clasificador a otro. Por ejemplo, la OIT instruyó que para CIUO 08 se codifique en 52 a las personas ocupadas en "Venta por catálogo", pero dado que el antiguo criterio para codificación con CIUO 88 consistía en que esos casos debían ser codificados con 91, el modelo se entrena con la información de que "venta por catálogo" debe ser codificado en 91, ante lo cual se requiere edición ex post del caso estimado con SVM, pues existe un nuevo criterio definido[13].

En concreto, esta fase consiste en un análisis posterior a la codificación automática con SVM, ya que, a pesar de la automatización del proceso, la "máquina entrenada" en el marco de la minería de texto, requiere el trabajo de analistas especializados en los clasificadores internacionales analizados, debido principalmente al dinamismo de las actividades económicas y el surgimiento de nuevas formas de trabajo dentro del mercado laboral. Si bien el sistema "aprende" y logra generar hallazgos inéditos, reconociendo patrones recurrentes en grandes volúmenes de textos, necesita que un analista nutra de información emergente de manera continua. Para ilustrar lo anterior, el caso de las plataformas tecnológicas, por ejemplo, es claro, pues casos en donde glosas describan actividades u ocupaciones asociadas a palabras no contenidas en la base de entrenamiento, por ejemplo "chofer Uber", requiere de análisis humano que asocie la palabra Uber al clasificador correspondiente.

Por otra parte, hay que considerar que las glosas clasificadas, son declaradas por un "informante idóneo", vale decir, cualquier persona que resida en el hogar de 15 años o más, lo que implica una serie de dificultades en cuanto al levantamiento y clasificación de la información. Por ejemplo, un informante puede "frasear" una misma ocupación de dos maneras distintas, lo que podría inducir a un error. En este contexto, es relevante mantener un análisis ex-post del comportamiento de casos seleccionados en base a los criterios derivados del análisis de los resultados del entrenamiento desarrollado durante el segundo semestre de 2018.

Finalmente, el uso de variables auxiliares para ayudar la clasificación es una tarea fundamental cuando estas no son posible de incorporar directamente en el entrenamiento, debido a las limitaciones propias del SVM. En este sentido, es necesario supervisar y editar también en base a

[12] Glosas especificadas en el clasificador en que su especificación, a todo evento corresponde ser clasificadas en un código específico. Por ejemplo, gerente general corresponde al subgrupo principal 11 de Directores ejecutivos y gerentes generales.

[13] Si bien en este caso se puede adoptar un método sistemático de edición, se sugiere que en estos casos se sostenga un seguimiento de los casos durante un período de tiempo.

los criterios que se establecen explícitamente en los clasificadores, tal como en el caso del CIUO 08, en donde la OIT releva, además de las tareas desempeñadas, la priorización de competencias para la clasificación de una determinada ocupación.

Un foco fundamental de este análisis en el desarrollo de la metodología ha sido la revisión de los casos dentro de los porcentajes de imprecisión, es decir, aquellos en que la codificación automática con SVM no coincide con el código asignado manualmente. Así pues, se deduce que el error debe tipificarse en dos categorías, uno, el error sistemático que hace referencia a un error en la aplicación de CA-SVM, y un segundo, correspondiente al error no sistemático, que en este caso correspondería a una mala codificación manual.

Para la determinación de casos a revisar se considerará de manera permanente:

- Casos críticos derivados de la matriz de transición (Ver sección 6 de Resultados).
    - Ejemplo: Casos que transitan entre Agricultura, Manufactura y Comercio, debido a sensibilidad de palabras clave; Venta, Producción, Exportación, etc.
- Casos entrantes a la muestra.
- Casos existentes en la muestra que cambian de código entre períodos de levantamiento (validación intertemporal).
- Casos cuya relación Sección/División en su clasificación no corresponde.
- Casos con probabilidad de clasificación menor a 0,8.
- Casos codificados en 999, que significa, casos que fueron clasificados como texto con información insuficiente, por lo que requiere uso de información auxiliar.

En términos operativos, esta fase determina el cierre de dos actividades del procesamiento de la ENE. Por una parte, al final del período de levantamiento y digitación de una submuestra[14] de la encuesta, se sintetizan las ediciones realizadas y se cargan en el sistema de información de la ENE para así permitir la construcción de la respectiva base de datos del trimestre móvil correspondiente y sus productos derivados (banco de datos ENE y publicación de indicadores en INE.stat). Por la otra parte, los casos analizados y editados son utilizados para retroalimentar el modelo que se deberá utilizar en los próximos períodos.

Fase 3: Retroalimentación

Como consecuencia del análisis ex-post, deriva la edición de la clasificación estimada con SVM, la que, en conjunto con las nuevas glosas y en particular de nuevas palabras detectadas en la última pesquisa, deben ser incorporados en el marco de entrenamiento que se utilizará en la codificación de los siguientes casos requeridos.

Esta actividad redunda en la actualización de los insumos R que utiliza el sistema compartido, derivados del proceso SVM, así como ediciones o nuevas reglas que puedan surgir, manteniendo

[14] Corresponde a la información de un mes específico del año.

consistencia con las recomendaciones de los organismos internacionales y las adaptaciones nacionales establecidas por el INE.

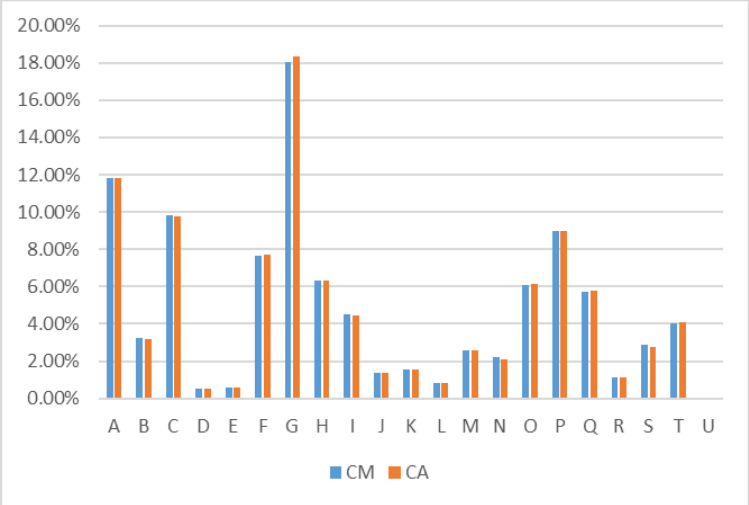
CAPITULO VI: RESULTADOS

En esta sección se analizan los resultados de los entrenamientos actualizados a 2018 que permitieron generar los modelos SVM e insumos requeridos para desarrollar el proceso de clasificación y codificación en la ENE a partir del trimestre febrero-abril de2019. También se describe el resultado de aplicar el proceso de análisis y edición descrito en la sección anterior a datos nuevos, es decir, fuera de la base de entrenamiento utilizada para determinar el modelo. Por último, se analiza si la introducción del método de clasificación automático tiene efecto significativo o no en las estimaciones oficiales derivadas con la encuesta.

6.1. Modelo SVM para CAENES

Los resultados de este entrenamiento arrojan que a nivel de sección (1° dígito), se alcanza un rendimiento promedio de 96,0% al contrastar lo codificado manualmente con lo automático en los casos de enero a diciembre 2018 utilizados para dicho entrenamiento. Este resultado, analizando la distribución sectorial de los casos, significa que el mayor efecto de la codificación automática se plasma en sector G de comercio, en donde, según la codificación manual los casos corresponderían a 18,1% del total de 184.267 registros que alimentan la base de entrenamiento 2018, al tiempo que, según la codificación automática, el mismo sector se compone de un total de 18,3%. En todos los demás sectores, las diferencias no superan el 0,1 punto porcentual.

Figura 5: Distribución % ocupados en la base de entrenamiento 2018, según CA y CM para CAENES a 1 dígito[15], enero-diciembre 2018.

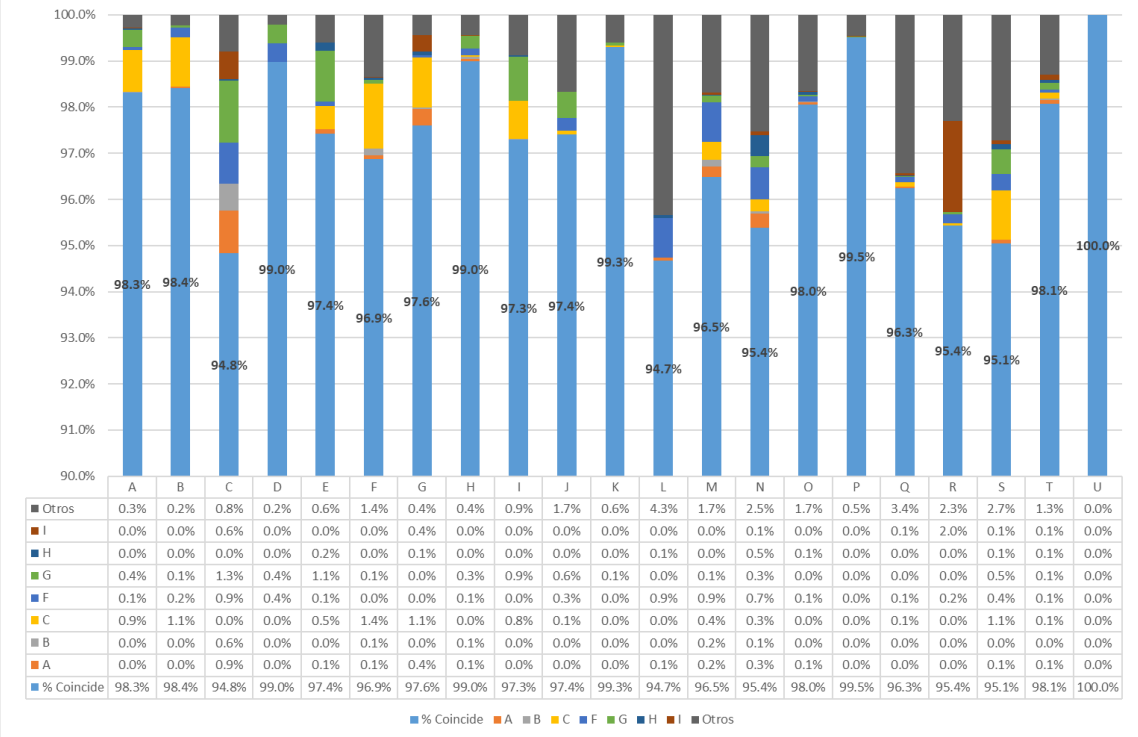


Fuente: Elaboración propia.

[15] Las secciones (1° dígito) del clasificador se encuentran especificadas en el anexo 1.

Con el propósito de comprender las diferencias entre ambos métodos, se procedió a analizar la composición de los casos que son codificados automáticamente, pues así se identifican los flujos intersectoriales que alimentan las secciones resultantes. El gráfico de la figura 6 permite dilucidar cuales son los sectores que alimentan la generación de cada sector clasificado automáticamente, así la barra de color azul, representa los casos que son clasificados en el mismo código con ambos métodos. En consecuencia, por ejemplo, se puede señalar que el sector T de organismos extraterritoriales se compone en un 100% por casos que de manera manual fueron codificados en el mismo código. Del mismo modo, se observa que uno de los sectores que tiene menor grado de coincidencia es el C de industria manufacturera, en donde, de las glosas que componen el sector C clasificado automáticamente, 94,8% de los casos fueron clasificados en el mismo sector cuando se hizo de manera manual. Siguiendo el análisis, se deduce también que el sector de Industria Manufacturera queda compuesto por actividades declaradas que manualmente fueron clasificadas en los sectores G de comercio (1,3%), F de construcción (0,9%) y A de agricultura (0,9%).

Figura 6: Composición de sectores económicos CA, según CM para CAENES a 1 dígito, enero-diciembre 2018.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Otros	0.3%	0.2%	0.8%	0.2%	0.6%	1.4%	0.4%	0.4%	0.9%	1.7%	0.6%	4.3%	1.7%	2.5%	1.7%	0.5%	3.4%	2.3%	2.7%	1.3%	0.0%
I	0.0%	0.0%	0.6%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.1%	2.0%	0.1%	0.1%	0.0%
H	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.5%	0.1%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%
G	0.4%	0.1%	1.3%	0.4%	1.1%	0.1%	0.0%	0.3%	0.9%	0.6%	0.1%	0.0%	0.1%	0.3%	0.0%	0.0%	0.0%	0.0%	0.5%	0.1%	0.0%
F	0.1%	0.2%	0.9%	0.4%	0.1%	0.0%	0.0%	0.1%	0.0%	0.3%	0.0%	0.9%	0.9%	0.7%	0.1%	0.0%	0.1%	0.2%	0.4%	0.1%	0.0%
C	0.9%	1.1%	0.0%	0.0%	0.5%	1.4%	1.1%	0.0%	0.8%	0.1%	0.0%	0.0%	0.4%	0.3%	0.0%	0.0%	0.1%	0.0%	1.1%	0.1%	0.0%
B	0.0%	0.0%	0.6%	0.0%	0.0%	0.1%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
A	0.0%	0.0%	0.9%	0.0%	0.1%	0.1%	0.4%	0.1%	0.0%	0.0%	0.0%	0.1%	0.2%	0.3%	0.1%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%
% Coincide	98.3%	98.4%	94.8%	99.0%	97.4%	96.9%	97.6%	99.0%	97.3%	97.4%	99.3%	94.7%	96.5%	95.4%	98.0%	99.5%	96.3%	95.4%	95.1%	98.1%	100.0%

Fuente: Elaboración propia.

La figura 6 deriva de la matriz de transición de CAENES, la cual refleja los casos codificados de manera manual en las filas y, los casos codificados automáticamente en las columnas. De ella deriva la composición de cada grupo clasificado automático, según el grupo al que correspondía

si fuera clasificado de manera manual. El análisis, a partir de técnicas de minería de textos, permitió una mayor comprensión de los flujos existentes entre actividades y de este modo se permitió elaborar una serie de criterios de casos a analizar una vez aplicado el modelo SVM. Con la ayuda de las nubes de palabras y el análisis crítico de casos individuales, fue posible discernir, por ejemplo, que en el flujo comercio-manufactura, el SVM tiende a considerar mayor cantidad de casos en comercio, principalmente, por casos que manualmente se clasificaban en manufactura debido a que se observa alta prevalencia de casos que en la glosa describen actividades de venta de pan, junto con abarrotes, bebidas, etc., lo cual induce a creer que corresponde más bien a actividades de almacén, más que de panadería y que por heterogeneidad de criterios, manualmente se clasificaban en manufactura por error. Por otra parte, el flujo inverso, permite identificar que existe alta prevalencia de casos que corresponden a barracas de madera que manualmente se clasificaban en comercio, al tiempo corresponde que sean clasificadas en manufactura, tal como ocurre con SVM. Ahora bien, la mayor incidencia se origina por la palabra venta, lo cual invita a concluir que es pertinente además sostener el seguimiento de estos casos en terreno para evitar el abuso de la palabra a la hora de describir una actividad económica. Similar situación ocurre en los casos que transitan entre construcción y manufactura, por lo que el modelo SVM enfrenta sensibilidades ante los pesos relativos de las palabras que se traducen en clasificaciones que, a veces corrigen los errores no muestrales de un proceso ejecutado de manera manual, pero también pueden caer en sesgos, lo que se ha denominado error sistemático para fines de este trabajo.

En definitiva, se sugiere sostener el análisis de resultados con SVM para efectos de clasificación automática de las glosas de la encuesta, a través de un análisis permanente a una muestra de los casos, focalizado en casos sensibles, tal que permita controlar, tanto los errores derivados de la codificación manual, denominados errores no sistemáticos, como también los errores sistemáticos del modelo.

Como consecuencia, se determinó un set de casos susceptibles de ser analizados de manera continua, tal como se muestra con el análisis de textos en los casos entre agricultura, industria manufacturera o comercio, o bien entre industria manufacturera o construcción.

Nube de palabras 5: Glosas de casos que transitan de Industria Manufacturera a Comercio.

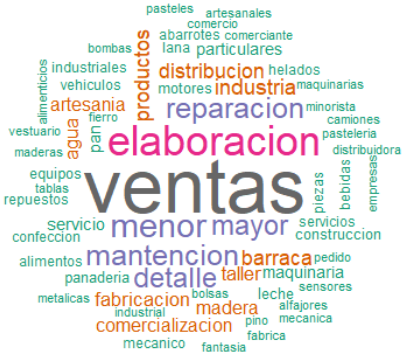
Fuente: Elaboración propia en base a la ENE 2018.

Tabla 6: Top 5 bigramas[16] de casos que transitan de Industria Manufacturera a Comercio

Ranking	Bigrama	Frecuencia
1	ventas menor	34
2	ventas mayor	30
3	ventas detalle	25
4	distribucion ventas	16
5	ventas pan	10

Fuente: Elaboración propia en base a la ENE 2018.

Nube de palabras 6: Glosas de casos que transitan de Comercio a Industria Manufacturera.



Fuente: Elaboración propia en base a la ENE 2018.

[16] Corresponde a la asociación de palabras con mayor grado de correlación entre sí, dentro de un conjunto específico de textos, Se utilizó la librería tidytext.

Tabla 7: Top 5 bigramas de casos que transitan de Comercio a Industria Manufacturera.

Ranking	Bigrama	Frecuencia
1	elaboracion ventas	11
2	ventas menor	11
3	reparacion mantencion	9
4	ventas detalle	7
5	barraca elaboracion	6

Fuente: Elaboración propia en base a la ENE 2018.

Nube de palabras 7: Glosas de casos que transitan de Industria Manufacturera a Construcción.

Fuente: Elaboración propia en base a la ENE 2018.

Tabla 8: Top 5 bigramas de casos que transitan de Industria Manufacturera a Construcción

Ranking	Bigrama	Frecuencia
1	estructuras metalicas	16
2	construccion estructuras	6
3	rejas protecciones	6
4	fabricacion instalacion	5
5	servicio construccion	5

Fuente: Elaboración propia en base a la ENE 2018.

Nube de palabras 8: Glosas de casos que transitan de Construcción a Industria Manufacturera.

Fuente: Elaboración propia en base a la ENE 2018.

Tabla 9: Top 5 bigramas de casos que transitan de Construcción a Industria Manufacturera.

Ranking	Bigrama	Frecuencia
1	estructuras metalicas	32
2	fabricacion estructuras	10
3	mantencion reparacion	7
4	metalicas galpones	5
5	reparacion mantencion	5

Fuente: Elaboración propia en base a la ENE 2018.

Para considerar la totalidad de las recomposiciones intersectoriales, a continuación, se presenta el resultado de una revisión exhaustiva realizada a los casos de la submuestra de febrero, en donde cuantifica el impacto del análisis ex-post en los resultados.

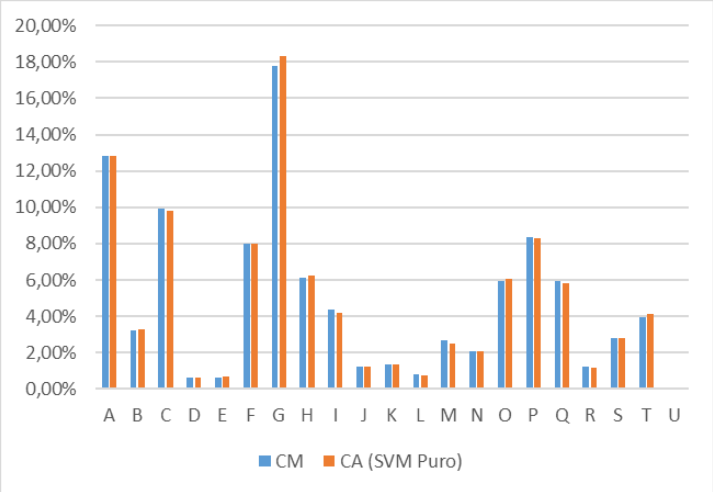
6.2. Resultados en datos nuevos con y sin revisión

El modelo determinado con data de entrenamiento 2018 fue aplicado a las glosas de la pregunta B14 pesquisadas en el trimestre móvil diciembre-febrero 2019, cuyos datos de enero y febrero corresponden a casos nuevos, es decir, casos fuera de la base de entrenamiento.

La ejecución de la fase de análisis y edición se llevó a cabo a partir de la ejecución del modelo SVM óptimo sobre los casos correspondientes a la submuestra de febrero 2019, a partir de lo cual se auditaron 1.160 casos correspondientes a la totalidad de casos que no coincidían entre ambos métodos. Lo anterior derivó en análisis de los casos y posteriores ediciones en base a los criterios definidos en el clasificador.

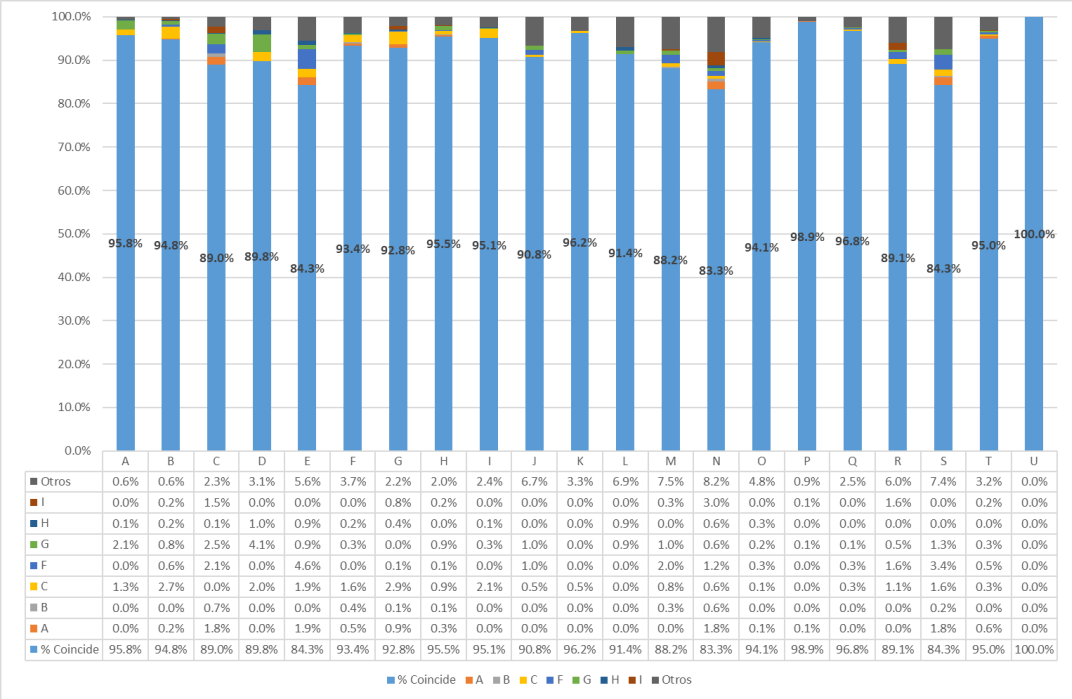
Para efectos de la descripción del ejercicio, a los resultados del modelo SVM sin auditoría, se le llama SVM Puro, mientras que los resultados auditados se denominan SVM Revisado. Se contrastan con los datos observados que corresponden a la Codificación Manual (CM), todo esto a nivel muestral.

Figura 7: Distribución % ocupados, según CA con SVM Puro y CM para CAENES a 1 dígito. Submuestra febrero 2019.



Fuente: Elaboración propia.

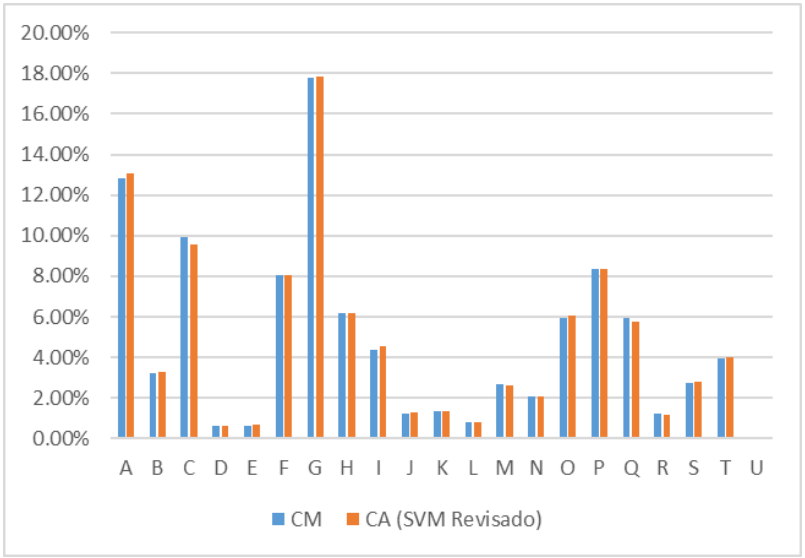
Figura 8: Composición de sectores económicos CA con SVM Puro, según CM para CAENES a 1 dígito. Submuestra febrero 2019.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Otros	0.6%	0.6%	2.3%	3.1%	5.6%	3.7%	2.2%	2.0%	2.4%	6.7%	3.3%	6.9%	7.5%	8.2%	4.8%	0.9%	2.5%	6.0%	7.4%	3.2%	0.0%
I	0.0%	0.2%	1.5%	0.0%	0.0%	0.0%	0.8%	0.2%	0.0%	0.0%	0.0%	0.0%	0.3%	3.0%	0.0%	0.1%	0.0%	1.6%	0.0%	0.2%	0.0%
H	0.1%	0.2%	0.1%	1.0%	0.9%	0.2%	0.4%	0.0%	0.1%	0.0%	0.0%	0.9%	0.0%	0.6%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
G	2.1%	0.8%	2.5%	4.1%	0.9%	0.3%	0.0%	0.9%	0.3%	1.0%	0.0%	0.9%	1.0%	0.6%	0.2%	0.1%	0.1%	0.5%	1.3%	0.3%	0.0%
F	0.0%	0.6%	2.1%	0.0%	4.6%	0.0%	0.1%	0.1%	0.0%	1.0%	0.0%	0.0%	2.0%	1.2%	0.3%	0.0%	0.3%	1.6%	3.4%	0.5%	0.0%
C	1.3%	2.7%	0.0%	2.0%	1.9%	1.6%	2.9%	0.9%	2.1%	0.5%	0.5%	0.0%	0.8%	0.6%	0.1%	0.0%	0.3%	1.1%	1.6%	0.3%	0.0%
B	0.0%	0.0%	0.7%	0.0%	0.0%	0.4%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.3%	0.6%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%
A	0.0%	0.2%	1.8%	0.0%	1.9%	0.5%	0.9%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	1.8%	0.1%	0.1%	0.0%	0.0%	1.8%	0.6%	0.0%
% Coincide	95.8%	94.8%	89.0%	89.8%	84.3%	93.4%	92.8%	95.5%	95.1%	90.8%	96.2%	91.4%	88.2%	83.3%	94.1%	98.9%	96.8%	89.1%	84.3%	95.0%	100.0%

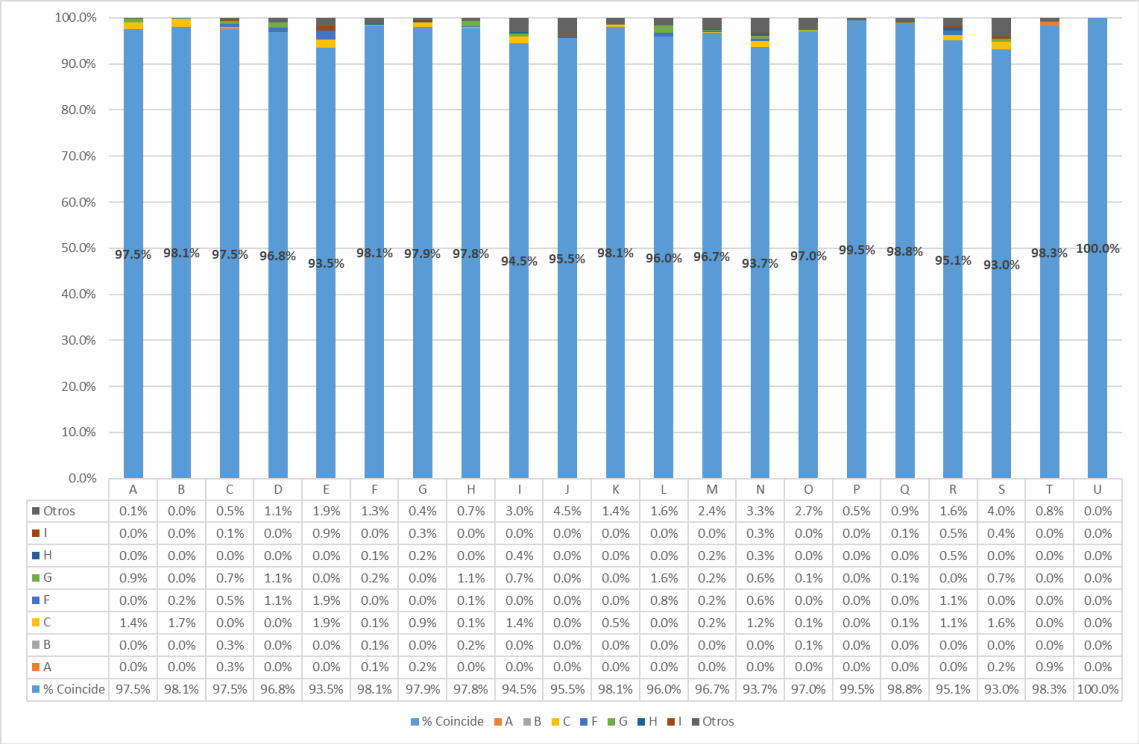
Fuente: Elaboración propia.

Figura 9: Distribución % ocupados, según CA con SVM Auditado y CM para CAENES a 1 dígito. Submuestra febrero 2019.



Fuente: Elaboración propia.

Figura 10: Composición de sectores económicos CA con SVM Auditado, según CM para CAENES a 1 dígito. Submuestra febrero 2019.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Otros	0.1%	0.0%	0.5%	1.1%	1.9%	1.3%	0.4%	0.7%	3.0%	4.5%	1.4%	1.6%	2.4%	3.3%	2.7%	0.5%	0.9%	1.6%	4.0%	0.8%	0.0%
I	0.0%	0.0%	0.1%	0.0%	0.9%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.1%	0.5%	0.4%	0.0%	0.0%
H	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.2%	0.0%	0.4%	0.0%	0.0%	0.0%	0.2%	0.3%	0.0%	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%
G	0.9%	0.0%	0.7%	1.1%	0.0%	0.2%	0.0%	1.1%	0.7%	0.0%	0.0%	1.6%	0.2%	0.6%	0.1%	0.0%	0.1%	0.0%	0.7%	0.0%	0.0%
F	0.0%	0.2%	0.5%	1.1%	1.9%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.8%	0.2%	0.6%	0.0%	0.0%	0.0%	1.1%	0.0%	0.0%	0.0%
C	1.4%	1.7%	0.0%	0.0%	1.9%	0.1%	0.9%	0.1%	1.4%	0.0%	0.5%	0.0%	0.2%	1.2%	0.1%	0.0%	0.1%	1.1%	1.6%	0.0%	0.0%
B	0.0%	0.0%	0.3%	0.0%	0.0%	0.1%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
A	0.0%	0.0%	0.3%	0.0%	0.0%	0.1%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.9%	0.0%
% Coincide	97.5%	98.1%	97.5%	96.8%	93.5%	98.1%	97.9%	97.8%	94.5%	95.5%	98.1%	96.0%	96.7%	93.7%	97.0%	99.5%	98.8%	95.1%	93.0%	98.3%	100.0%

Fuente: Elaboración propia.

Como resultado del ejercicio, se muestra que:

- Las predicciones sobre datos nuevos reducen el nivel de rendimiento promedio alcanzado en la base de entrenamiento.
- El análisis y edición ex-post impacta positivamente los resultados de lo codificado con el modelo SVM.

Ahora bien, a pesar de que la revisión mejora considerablemente los resultados, comparando con lo codificado manualmente, se observa que no siempre se igualan idénticamente los niveles observados con ambos métodos.

Como se ha señalado anteriormente, ambos métodos están sujetos a errores. Los errores de la codificación automática se reconocen como errores sistemáticos, los cuales pueden ser abordados a través del análisis especializado. En este sentido, cabe señalar que las diferencias que se observan corresponden siempre al efecto de recomposición entre ambas técnicas, es decir, clasificaciones que con metodología automática derivan en códigos diferentes a lo determinado de manera manual. Así entonces, es posible considerar que casos como la recomposición agricultura-manufactura está determinada principalmente por actividades integradas, pero que a partir de

información auxiliar se permite precisar, o aproximar de mejor manera, que corresponde a labores de establecimientos agrícolas.

En este sentido, se podría concluir entonces que los casos en que el método automático, post revisión, no logra igualar el código manual, corresponde al hecho de que los casos codificados manualmente tenían algún tipo de error no muestral propio del proceso.

6.3. Modelo SVM para CIUO 08.CL

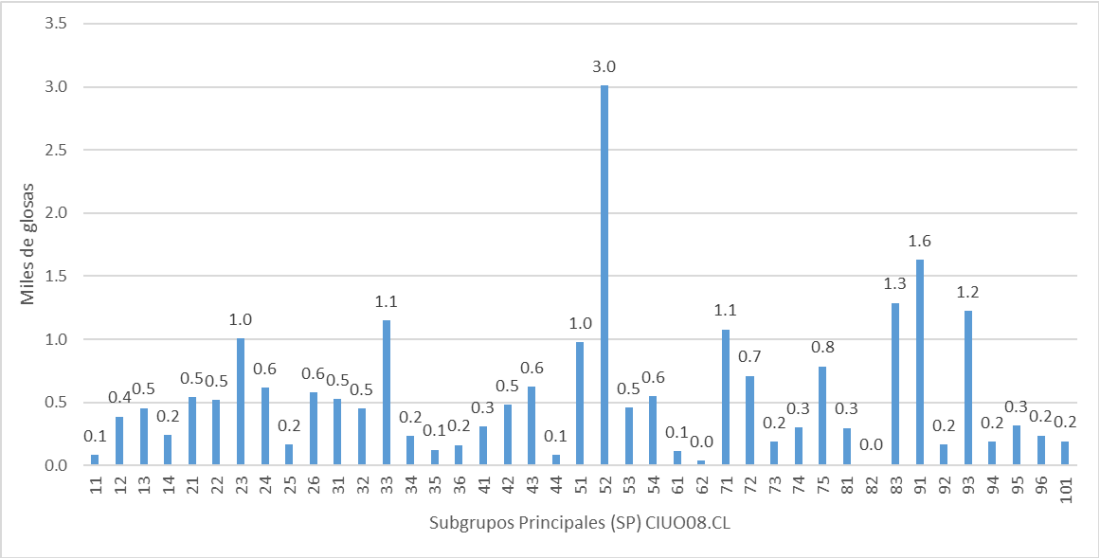
Para la determinación de modelos SVM que permiten clasificar las preguntas de ocupación, se construyó una base de entrenamiento basado en la totalidad de las ocupaciones pesquisadas en la octava Encuesta de Presupuestos Familiares (EPF VIII), equivalente a 22.539 glosas, las cuales fueron codificadas manualmente en una versión preliminar de CIUO 08.CL durante el año 2018[17]. Esto se debió a que la ENE ha sido codificada manualmente solo en CIUO 88, por lo que era imposible entrenar con sus propios datos, toda vez que el nuevo clasificador contiene transiciones entre códigos propias de la actualización, como por ejemplo, la transición de los guardias de seguridad del gran grupo 9 de trabajadores no calificados al gran grupo 5 de trabajadores de los servicios, entre otras transiciones.

Considerando que la EPF se levanta solo en áreas urbanas, se produce un efecto de escasez de glosas si se utiliza solo casos de EPF para clasificar las glosas de la ENE, pues esta última al ser de cobertura nacional, pesquisa ocupaciones que no aparecen en la base de entrenamiento. Para subsanar esta carencia, se agregó un conjunto de 805 glosas de ocupación que en la ENE fueron codificadas manualmente en CIUO 88, pero se recodificaron en CIUO 08.CL. Los casos se concentraron principalmente en el gran grupo 6 de Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros, más casos que actualmente son clasificados como peones agropecuarios, forestales, pesqueros (Subgrupo Primario 92), así como también casos del Subgrupo Primario 82, específicamente ocupaciones de “Embaladores” donde también se determinó que era necesario alimentar la base de entrenamiento con un mayor volumen de casos.

La figura 11 refleja la distribución, en miles, de las 22.539 glosas de EPF por subgrupo principal del clasificador. La tabla 10 detalla la cantidad de casos adicionados a la base de entrenamiento y, para cuantificar el efecto en la distribución porcentual de casos, se presentan los gráficos de las figuras 12 y 13, donde se puede apreciar que se triplicó el peso relativo de los subgrupos principales 61 y 92 y, se duplicó el peso relativo del resto de los subgrupos que se alimentaron de información ENE. Como consecuencia, se beneficia significativamente el resultado sobre la ENE, pues en los experimentos que permitieron determinar el modelo óptimo pudo observarse que, si no se consideraban glosas como las agregadas a la base de entrenamiento pura, no era posible clasificar ningún caso en los códigos incorporados, lo cual es una necesidad real para la ENE.

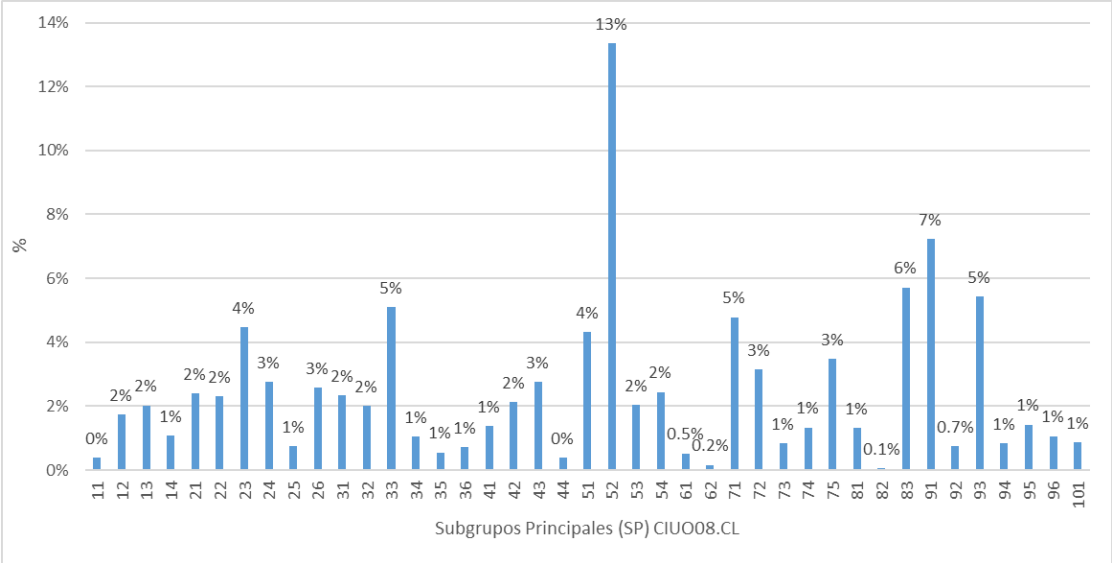
[17] Metodología de codificación EPF VIII. https://www.ine.cl/docs/default-source/ingresos-y-gastos/epf/viii-epf/documentacion/metodolog%C3%ADa-viii-epf.pdf?sfvrsn=4

Figura 11: Distribución de glosas EPF (en miles), según CIUO 08.CL.



Fuente: Elaboración propia a partir de EPF VIII

Figura 12: Distribución % glosas EPF VIII, según CIUO 08.CL



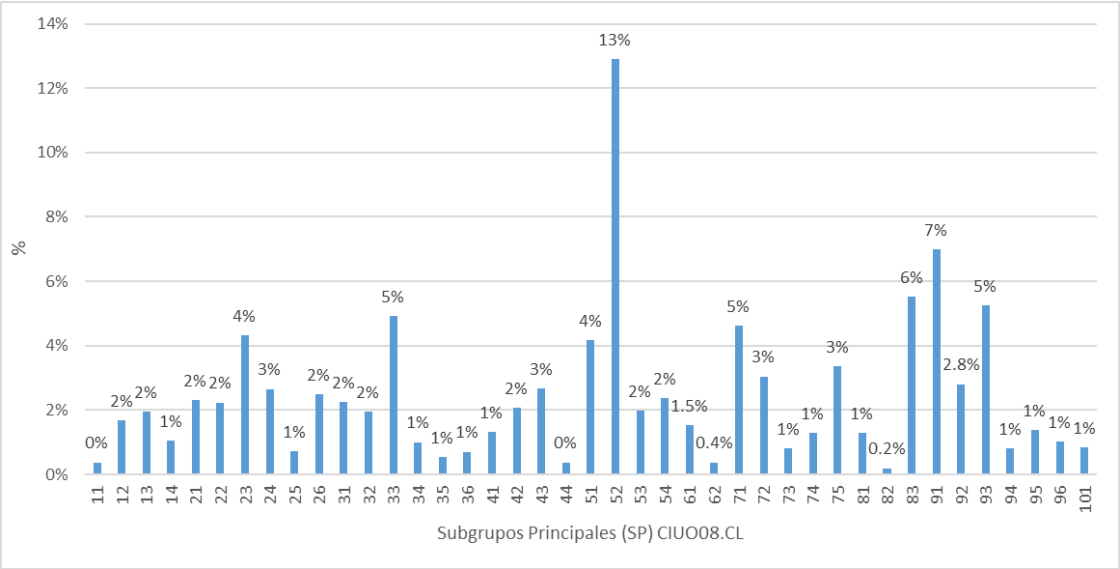
Fuente: Elaboración propia a partir de EPF VIII

Tabla 10: Distribución de glosas ENE recodificadas a CIUO 08.CL

SP	Glosas
61	238
62	50
82	31
92	486
Total	**805**

Fuente: Elaboración propia a partir de casos ENE 2017.

Figura 13: Distribución % glosas set ampliado (EPF VIII + ENE)



Fuente: todos los gráficos y tablas son elaboración propia, a partir de EPF VIII y ENE 2017.

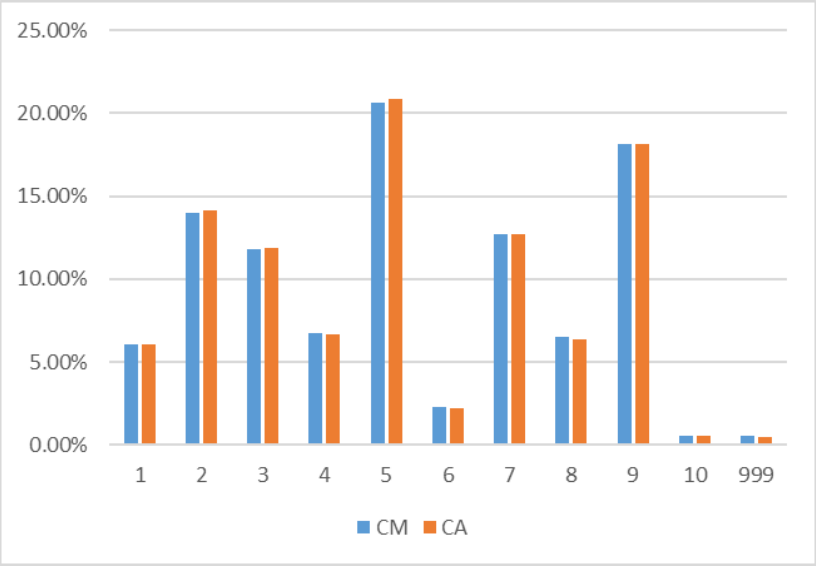
Una vez desarrollado este marco de entrenamiento idóneo, se procedió con la metodología SVM determinada y se entrenó con el 80% de los casos, para luego probar con el 20% restante, encontrándose rendimientos en torno a 90%-95%. Así, se determinó un modelo SVM óptimo para clasificar ocupación con CIUO 08.CL y se aplicó sobre los datos de la ENE del año 2017. Posteriormente, los resultados del modelo sobre el entrenamiento y sobre los datos de la ENE fueron compartidos a la sección de nomenclaturas del INE, en donde se procedió a auditar una muestra de cuatro mil casos, focalizada en aquellos casos donde, en la base de entrenamiento, no coincidía el código de ambos métodos de codificación y, en el caso de la ENE, se focalizó en casos críticos de codificación que requieren información auxiliar tales como el nivel de competencias. Esto último debido a que en la ENE, la actualización del clasificador impide considerar equivalente los códigos entre el CIUO 88 y el CIUO 08.CL, en este sentido, se hizo fundamental comprender la dinámica entre las glosas declaradas y los resultados del modelo siguiendo las directrices del clasificador, lo que, junto con la experiencia de profesionales del INE en materias de clasificación, generó una sinergia para el proceso global de clasificación y codificación que se determinó era

fundamental para la aplicación de cualquier método automático que clasifique variables de interés para las estadísticas del trabajo.

Posteriormente, se introdujeron los casos auditados en un nuevo entrenamiento, resultando que el modelo coincidió en el 99,7% de estos casos. En consecuencia, se obtuvo un modelo SVM para el nuevo clasificador con un rendimiento promedio que alcanza al 94,9% de rendimiento a 1 dígito y de un 98,8% a nivel de 2 dígitos.

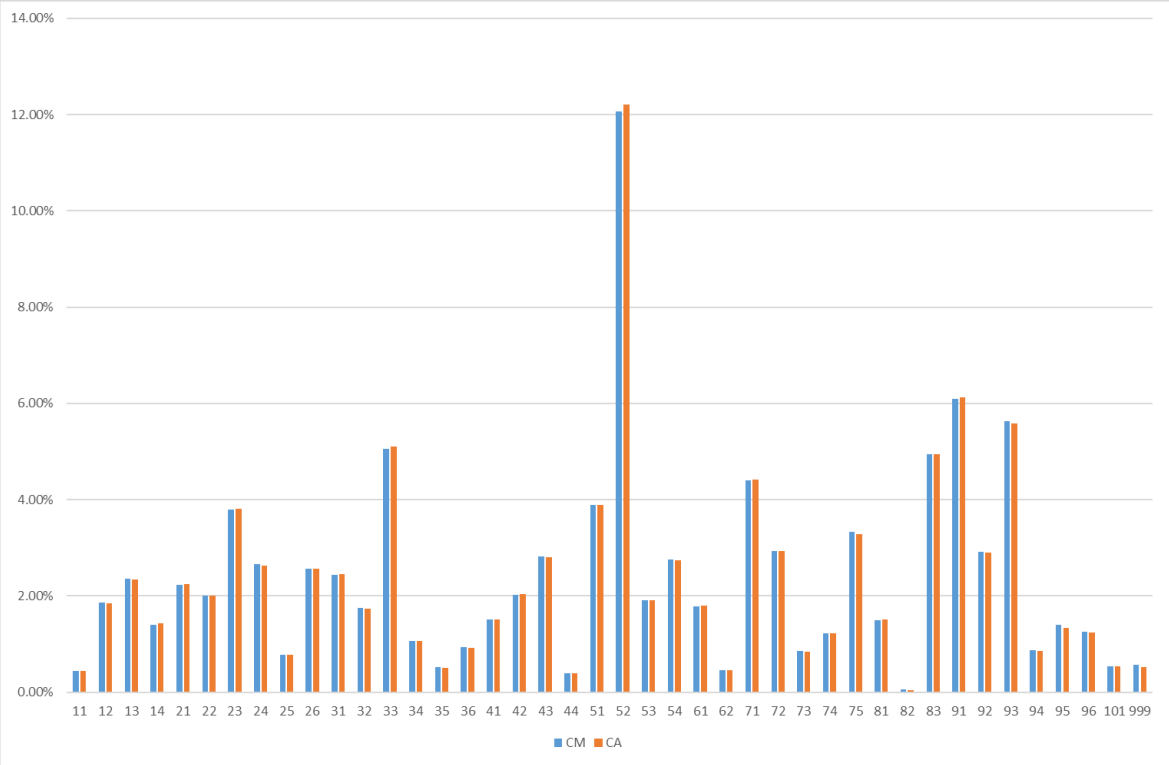
Una primera medida visual del rendimiento del modelo se puede apreciar al contrastar los casos en cada grupo del calificador, según cada técnica de codificación, en donde se ve reflejado de muy buena manera que no se observan grandes cambios en los casos que correspondería clasificar en cada grupo (Ver figuras 14 y 15).

Figura 14: Distribución % ocupados en la base de entrenamiento, según CA y CM para CIUO 08.CL a 1 dígito. Enero-diciembre 2018.



Fuente: Elaboración propia en base a glosas EPF VIII y ENE.

Figura 15: Distribución % ocupados en la base de entrenamiento, según CA y CM para CIUO 08.CL a 2 dígitos.

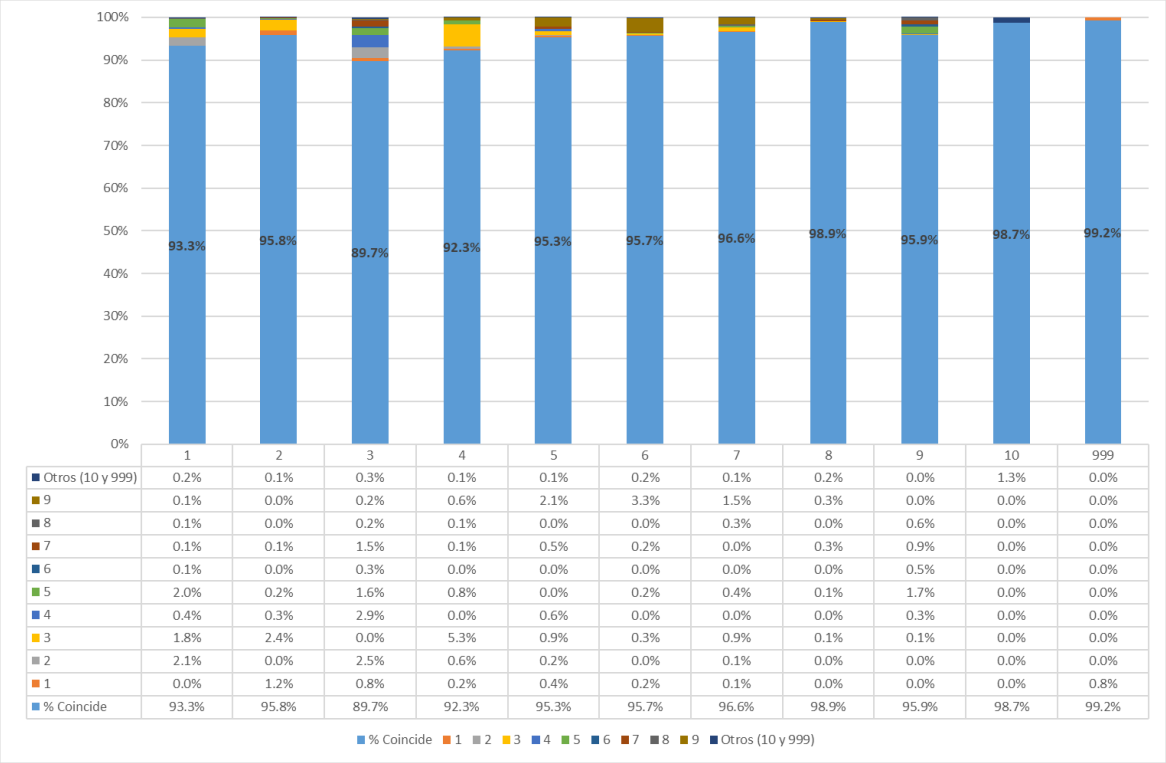


Fuente: Elaboración propia en base a glosas EPF VIII y ENE.

Una segunda medida visual consiste en verificar la transición de los casos que, codificados de manera manual, pasan a quedar en un diferente grupo al ser clasificados de manera automática. Para esto, se utiliza la matriz de transiciones, a partir de la cual se elaboró la figura 16 que refleja la composición de cada grupo clasificado automáticamente, según el grupo al que correspondía si fuera clasificado de manera manual. La barra de color azul refleja el porcentaje de casos en donde los casos coinciden en el código asignado con ambos métodos. De esta medida se desprende, por ejemplo, que las mayores concentraciones de casos que cambian de código, según la técnica, se presentan en los casos que codificados de manera manual son clasificados en el grupo 3 de técnicos, no obstante, el modelo de clasificación automática determina que deben ser clasificados en el grupo 4 de trabajadores de oficina. Le siguen los casos que fueron codificados manualmente en el grupo 9 de trabajadores no calificados, sin embargo, el modelo automático determina que deben ser clasificados en el sector agrícola.

Los factores que determinaron las transiciones en los códigos no son muy distintos a los presentados en la sección anterior para el caso de actividad económica.

Figura 16: Composición de grupos ocupacionales CA, según CM, para CIUO 08.CL a 1 dígito.



	1	2	3	4	5	6	7	8	9	10	999
Otros (10 y 999)	0.2%	0.1%	0.3%	0.1%	0.1%	0.2%	0.1%	0.2%	0.0%	1.3%	0.0%
9	0.1%	0.0%	0.2%	0.6%	2.1%	3.3%	1.5%	0.3%	0.0%	0.0%	0.0%
8	0.1%	0.0%	0.2%	0.1%	0.0%	0.0%	0.3%	0.0%	0.6%	0.0%	0.0%
7	0.1%	0.1%	1.5%	0.1%	0.5%	0.2%	0.0%	0.3%	0.9%	0.0%	0.0%
6	0.1%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	0.0%
5	2.0%	0.2%	1.6%	0.8%	0.0%	0.2%	0.4%	0.1%	1.7%	0.0%	0.0%
4	0.4%	0.3%	2.9%	0.0%	0.6%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%
3	1.8%	2.4%	0.0%	5.3%	0.9%	0.3%	0.9%	0.1%	0.1%	0.0%	0.0%
2	2.1%	0.0%	2.5%	0.6%	0.2%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%
1	0.0%	1.2%	0.8%	0.2%	0.4%	0.2%	0.1%	0.0%	0.0%	0.0%	0.8%
% Coincide	93.3%	95.8%	89.7%	92.3%	95.3%	95.7%	96.6%	98.9%	95.9%	98.7%	99.2%

Fuente: Elaboración propia en base a glosas EPF VIII y ENE.

Para la comprensión de las transiciones que ocurren al cambiar al método automático, es fundamental señalar la relación existente entre la clasificación de ocupación y las competencias para desempeñarse como tal, así como señala OIT en el clasificador. Este tópico se intentó abordar en forma de texto para el entrenamiento del modelo, a partir de la descripción del nivel educacional alcanzado. Sin embargo, para efectos de clasificación con el modelo SVM, no se observaron diferencias significativas entre utilizar o no dicha información.

Ahora bien, dada su relevancia y pertinencia conceptual, el vínculo de las competencias con la ocupación es uno de los focos centrales del análisis ex-post, las cuales son implementadas en el sistema a través de ediciones sistemáticas o puntuales que redundan en la edición de los casos correspondientes, tanto para la publicación mensual de los datos, así como para el reentrenamiento del modelo con fines de realizar su permanente y debida actualización.

Tabla 11: Matriz de transición entre CM y CA, CIUO 08.CL a 1 dígito.

		Codificación Automática (CA)											Total
		1	2	3	4	5	6	7	8	9	10	999	
Codificación Manual (CM)	1	1,575	45	27	4	24	1	4	-	-	-	1	1,681
	2	35	3,741	82	11	10	-	3	-	-	-	-	3,882
	3	31	95	2,944	98	51	2	31	2	4	-	-	3,258
	4	7	12	94	1,712	32	-	-	-	13	-	-	1,870
	5	33	6	52	15	5,502	1	13	2	85	-	-	5,709
	6	1	-	9	-	2	583	1	-	27	-	-	623
	7	1	5	49	1	26	1	3,394	6	44	-	-	3,527
	8	1	-	8	1	2	-	12	1,746	32	-	-	1,802
	9	1	-	6	11	121	20	54	6	4,810	-	-	5,029
	10	-	-	1	1	-	-	-	-	-	149	-	151
	999	3	2	9	-	3	1	3	3	2	2	128	156
Total		1,688	3,906	3,281	1,854	5,773	609	3,515	1,765	5,017	151	129	27,688

Fuente: Elaboración propia en base a glosas EPF VIII y ENE.

El análisis derivado de la matriz de transición de CIUO 08.CL, a nivel de 2 dígitos es una tarea propia del proceso de revisión de los resultados del modelo estimado, pero también para determinar los criterios con que se seleccionarán los casos que deberán ser revisados por el equipo de expertos en clasificadores, y de este modo focalizar la necesidad de mejoramiento continuo del proceso, una vez puesto en marcha el método automático.

CAPITULO VII: CONCLUSIONES

Se desarrolló un Sistema de Clasificación y Codificación Automática, a partir de los textos observados en la ENE durante 2018 y EPF VIII (2016 - 2017). De este proceso derivaron modelos SVM para los clasificadores requeridos por la ENE, a saber, CIUO-08.CL y CAENES. Su desarrollo fue posible gracias a la determinación de un modelo óptimo en el trabajo de Guerrero y Cabezas (2019) que utilizó data de la ENE observada durante abril de 2015 y diciembre de 2017.

Específicamente, la codificación de los casos pesquisados a partir del trimestre febrero-abril de 2019, que se basan en CAENES, utilizan modelos de clasificación automática de Support Vector Machine (SVM), determinados a partir de los entrenamientos con datos de la población ocupada de la ENE del año 2018. Para la codificación del clasificador CIUO 08.CL, se aplicó misma metodología de clasificación SVM y se utilizaron las ocupaciones clasificadas por la EPF VIII en dicho clasificador, más glosas de la ENE recodificadas de CIUO 88 a CIUO 08.CL. En particular, glosas clasificadas en los subgrupos principales de ocupaciones agrícolas, debido que la EPF recoge data en zonas urbanas y la ENE en territorio urbano y rural, pues se requiere contar con glosas representativas de las ocupaciones a nivel nacional.

El método desarrollado se basa en grandes tres pilares, a saber, clasificadores internacionales, Metodología SVM y, análisis ex-post de los resultados del modelo. De este modo, se permite

disponer de un proceso de clasificación y codificación global, tal que aborde específicamente los textos de la ENE referidos a “Ocupación, labor u oficio” y “Actividad económica” que declara la población objetivo en sus respectivas entrevistas. Para la ejecución del proceso de aplicación de los modelos basados en SVM, se necesitan los textos que requieren código en formato de tabla para su lectura en R Studio y posterior aplicación de los algoritmos desarrollados en la misma plataforma. Una vez ejecutada la metodología de aprendizaje de máquina, se procede con el análisis de los resultados por parte de los expertos en clasificadores, cuyo propósito consiste en realizar las ediciones correspondientes cuando se ameriten y así garantizar la calidad del proceso a partir de resultados más precisos y mejores bases de entrenamiento para la actualización de los modelos.

El sistema desarrollado se ejecuta en 4 fases. La primera de preparación de los modelos SVM con data previamente codificada; una segunda fase de ejecución del modelo SVM correspondiente; una tercera fase de análisis y edición de la codificación con SVM, la cual es el insumo final para la última fase de retroalimentación, cuyo foco es la actualización del modelo SVM a utilizar en la siguiente codificación requerida.

La fase de retroalimentación representa el valor agregado al proceso de clasificación automática con SVM. Permite corregir errores no sistemáticos (originados por la codificación manual), así como errores sistemáticos derivados de la sensibilidad a las palabras de todo método automático de clasificación de textos en estadísticas referidas a fenómenos laborales. Su metodología se desarrolló a partir de los análisis desarrollados durante este trabajo y se basaron en contrastar los casos codificados manualmente con los codificados de manera automática, a partir de técnicas de minería de textos. De este modo, se pudo abordar grandes volúmenes de casos y, se determinaron criterios específicos de análisis ex-post, los cuales son aplicados durante los últimos diez días de cada mes de digitación de casos. Como consecuencia de este análisis, derivan posibles ediciones a la clasificación de SVM, las que son adoptadas para la publicación de los datos, así como para la retroalimentación a los modelos SVM, a través de bases de entrenamiento permanentemente actualizadas y mejoradas.

Los modelos SVM estimados, al compararse con la data de entrenamiento, arrojan resultados entre 94,5%-98,8%, según el clasificador. Al aplicar estos modelos a nuevos textos que requieren codificación, pero complementado con análisis con fines de control de calidad, el resultado mejora hasta 4 puntos porcentuales, lo que significa pasar de un indicador de rendimiento de 93,4% a 97,4%.

Como consecuencia de este ejercicio, se podría señalar que el error de estimación, medido como el porcentaje de casos que no coinciden, en una fase inicial que no contempla supervisión, alcanzaría a 6,6%. No obstante, debido a la dinámica propia de un proceso de codificación manual, en que se observa alta heterogeneidad en la aplicación de criterios, es posible presumir que solo una parte de ese porcentaje corresponde a error, ya que la otra, corresponde a casos que el método corrige. En este sentido, considerando la revisión con fines de control de calidad del modelo, los casos codificados automáticamente que no coinciden se reducen a 2,6%, los cuales podría presumirse corresponde a errores no sistemáticos, considerados como casos que corresponde ser recodificados

en base al modelo SVM, pues este corrige el código asignado de manera manual. En consecuencia, los cuatro puntos porcentuales corregidos corresponderían al error no sistemático del período.

El desarrollo de este trabajo hizo comprender la necesidad de identificar claramente la dinámica entre las glosas declaradas y los resultados del modelo siguiendo las directrices del clasificador, lo que, junto con la experiencia de profesionales del INE en materias de clasificación, generó una sinergia para el proceso global de clasificación y codificación que se determinó era fundamental para la aplicación de cualquier método automático que clasifique variables de interés para las estadísticas del trabajo.

Finalmente, en términos de proceso, la nueva metodología sustituye un número importante de horas de trabajo al mes de dedicación exclusiva al acto de codificar las cinco preguntas de la ENE, por uno que las clasifica automáticamente en minutos y que permite focalizar el trabajo en revisiones posteriores por parte de analistas expertos en clasificadores que permiten asegurar la calidad de la codificación con fines de publicación coyuntural, así como de mantener una base de entrenamiento debidamente actualizada a las dinámicas propias del mercado laboral. A su vez, este proceso permite la adopción de nuevos clasificadores de manera más rápida, como es el caso del Clasificador Internacional Uniforme de Ocupaciones 2008 adaptado a Chile y que comenzará su publicación oficial a contar del trimestre febrero-marzo 2019.

BIBLIOGRAFÍA

CAENES, Clasificador de Actividades Económicas Nacional para Encuestas Sociodemográficas, INE. http://historico.ine.cl/canales/chile_estadistico/mercado_del_trabajo/empleo/metodologia/pdf/caenes.pdf

CIUO 08.CL, Clasificador Chileno de Ocupaciones, INE. https://www.ine.cl/docs/default-source/publicaciones/2018/ciuo-08.cl-clasificador-chileno-de-ocupaciones.pdf?sfvrsn=4

Contreras Barrera, Marcial (2014): "Minería de texto: una visión actual" Biblioteca Universitaria, vol. 17, núm. 2, julio-diciembre, 2014, pp. 129-138.

Guerrero J y Cabezas J (2019), "Clasificación automática de textos utilizando técnicas de text mining: Aplicación a las glosas de la Encuesta Nacional de Empleo (ENE)", INE. https://www.ine.cl/docs/default-source/documentos-de-trabajo/metodologicos/clasificacion-automatica-de-textos-utilizando-tecnicas-de-text-mining-aplicacion-a-las-glosas-de-la-encuesta-nacional-de-empleo.pdf?sfvrsn=0

Nashif, Shadman & Rakib Raihan, Md & Rasedul Islam, Md & Imam, Mohammad. (2018). Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World Journal of Engineering and Technology. 06. 854-873. 10.4236/wjet.2018.64057.

Sánchez, D., Martín-Bautista, M (2006). "Un enfoque deductivo para la minería de texto" [en línea].

ANEXOS

Secciones CAENES

SECCIÓN	DESCRIPCIÓN
A	Agricultura, ganadería, silvicultura y pesca
B	Explotación de minas y canteras
C	Industrias manufactureras
D	Suministro de electricidad, gas, vapor y aire acondicionado
E	Suministro de agua
F	Construcción
G	Comercio al por mayor y al por menor
H	Transporte y almacenamiento
I	Actividades de alojamiento y de servicio de comidas
J	Información y comunicaciones
K	Actividades financieras y de seguros
L	Actividades inmobiliarias
M	Actividades profesionales, científicas y técnicas
N	Actividades de servicios administrativos y de apoyo
O	Administración pública y defensa
P	Enseñanza
Q	Actividades de atención de la salud humana y de asistencia social
R	Actividades artísticas, de entretenimiento y recreativas
S	Otras actividades de servicios
T	Actividades de los hogares como empleadores
U	Actividades de organizaciones y órganos extraterritoriales

Subgrupos principales CIUO 08.CL

Subgrupo	Descripción
11	Miembros del poder ejecutivo y legislativo, personal directivo de la administración pública y de otras organizaciones sociales y/o políticas, directores ejecutivos y gerentes generales
12	Directores y gerentes administrativos y de servicios comerciales
13	Directores, gerentes y administradores de producción y operaciones
14	Directores, gerentes y administradores de hoteles, restaurantes, comercios y de otros servicios
21	Profesionales de las ciencias y de la ingeniería
22	Profesionales de la salud
23	Profesionales de la educación
24	Profesionales de negocios y administración
25	Profesionales de tecnología de la información y las comunicaciones
26	Profesionales en derecho, ciencias sociales y culturales
31	Técnicos de las ciencias y la ingeniería
32	Técnicos de la salud
33	Técnicos en operaciones financieras y administrativas
34	Técnicos de servicios jurídicos, sociales, deportivos y culturales
35	Técnicos de la tecnología de la información y las comunicaciones
36	Técnicos en educación
41	Oficinistas
42	Empleados en trato directo con el público
43	Auxiliares y ayudantes de registros contables y encargados del registro de materiales
44	Otro personal de apoyo administrativo
51	Trabajadores de los servicios a las personas
52	Vendedores
53	Trabajadores de los cuidados personales
54	Personal de los servicios de protección y seguridad
61	Agricultores y trabajadores calificados de explotaciones agropecuarias cuya producción se destina al mercado
62	Trabajadores forestales calificados, pescadores y cazadores cuya producción se destina al mercado
63	Trabajadores agropecuarios, pescadores, cazadores y recolectores de subsistencia
71	Operarios de la construcción (no incluye electricistas)
72	Operarios de la metalurgia y operarios de máquinas herramientas; mecánicos de vehículos, maquinarias, aviones y bicicletas
73	Artesanos y operarios de las artes gráficas
74	Trabajadores especializados en electricidad y electrónica
75	Operarios de procesamiento de alimentos, de la confección, ebanistas y otros oficios
81	Operadores de instalaciones fijas y máquinas
82	Ensambladores
83	Conductores de vehículos y operadores de equipos pesados y móviles
91	Auxiliares de aseo y trabajadores de casa particular
92	Obreros agropecuarios, pesqueros y forestales
93	Obreros de la minería, la construcción, la industria manufacturera y el transporte
94	Cocineros de comida rápida y ayudantes de cocina
95	Trabajadores ambulantes de servicios y vendedores ambulantes (excluyendo comida de consumo inmediato)
96	Recolectores de desechos y otras ocupaciones elementales
101	Otros no identificados