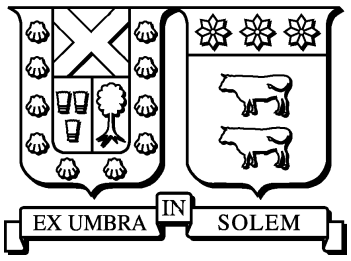
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



"APLICACIÓN DE TÉCNICAS DE *MACHINE LEARNING* PARA PREDECIR EL TAMAÑO DE INCENDIOS FORESTALES"

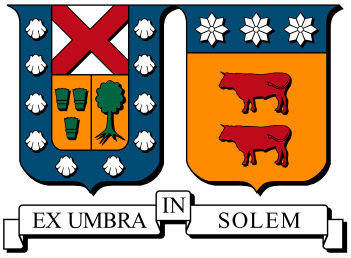
MATÍAS FELIPE CAMPOS SANTELICES

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: JOSÉ LUIS MARTÍ

NOVIEMBRE 2017

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"APLICACIÓN DE TÉCNICAS DE *MACHINE LEARNING* PARA PREDECIR EL TAMAÑO DE INCENDIOS FORESTALES"

MATÍAS FELIPE CAMPOS SANTELICES

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: JOSÉ LUIS MARTÍ
PROFESOR CORREFERENTE: RICARDO ÑANCULEF

NOVIEMBRE 2017

Agradecimientos

Agradezco a cada una de las personas que estuvieron presentes en el desarrollo de esta memoria, aportando con su alegría y palabras de aliento cuando fue necesario. En primer lugar agradezco al profesor guía José Luis Martí, quién me guío tanto en el planteamiento del problema de investigación como en su desarrollo. . Finalmente, también agradezco a mi familia y amigos, quienes me dieron sus palabras de aliento y optimismo para continuar con este trabajo, en especial a mis padres David y Carmen Gloria, y mis hermanos Tomás y María Paz.

Resumen

El control y contención de incendios forestales se ha convertido en una temática cada vez más importante alrededor del mundo, no tan solo por el daño que éstos causan al hábitat del ser humano, sino que también por el dinero invertido en su combate. Es por esto, que contar con sistemas que apoyen la toma de decisiones para afrontar estos eventos naturales es fundamental para asignar recursos de forma eficiente. El presente trabajo propone un modelo predictivo del tamaño de los incendios forestales utilizando técnicas de *machine learning*, en el cual mediante el uso del proceso de minería de datos CRISP-DM, se proponen dos acercamientos para resolver dicha tarea: primeramente prediciendo esta área de forma cuantitativa, aplicando un modelo de regresión; y por otra parte, cuantitativamente prediciendo su tamaño a través de etiquetas. En particular se evaluará el desempeño de las máquinas de soporte vectorial, evaluando las ventajas y desventajas de ambos acercamientos y comparando los resultados obtenidos con estudios similares.

Palabras Claves: Incendios Forestales, Máquinas de Soporte Vectorial, CRISP-DM, Clustering, Machine Learning.

Abstract

Controlling and containing forest fires has become an important topic of discussion around the world due to not only the great damaged they cause to human habitat but also the amounts of financial resources invested in their extinction. This is where lies the importance of having systems that support decision making on how to efficiently allocate resources to fight them. This present paper proposes a predictive modeling for the magnitude of forest fires using machine learning techniques. Through the use of CRISP-DM data mining process two proposal are presented. First, a quantitative prediction by applying a regression model, and second, quantitatively predicting its size through tags. In particular, the performance of Support Vector Machines will be assessed, evaluating the advantages and disadvantages of both approaches and comparing the results obtained with similar studies.

Keywords: Forest Fires, Support Vector Machines, CRISP-DM, Clustering, Machine Learning.

Índice de Contenidos

Índice de cuadros

Índice de figuras

Introducción

Los incendios forestales son un evento natural, que en su mayoría son originados por la acción del hombre y en menor parte, por rayos originados en tormentas eléctricas. Éstos, conforman parte del ciclo natural de la vida de los bosques, aumentando la diversidad de las especies y permitiendo el desarrollo de nueva flora y fauna. Sin embargo, si no se controlan a tiempo, éstos pueden afectar negativamente el ecosistema del ser humano.

El combate y control de los incendios forestales en diversos países se encuentra a cargo de entidades gubernamentales, por ejemplo en Canadá se encuentra el *Natural Resource Canada*, en Estados Unidos el *U.S Forest Service* y en Chile la Corporación Nacional Forestal (CONAF). Cada una de estas entidades ha implementado diferentes sistemas computacionales, los cuales entregan información que apoya el proceso de toma de decisiones y asignación de recursos.

En esta memoria, se desarrolla una propuesta que se enfoca en la predicción del área quemada por los incendios forestales a través del uso de las Máquinas de Soporte Vectorial. Siguiendo el proceso de minería de datos CRISP-DM, se evaluarán dos modelos para resolver esta problemática, uno de regresión y otro de clasificación, con el objetivo de evaluar su desempeño y comparar los resultados obtenidos con los registrados en estudios similares, contribuyendo así con un nuevo estudio en el área.

La estructura de la memoria se compone de cuatro capítulos, en el capítulo 1 se describe la repercusión de los incendios forestales en el medio ambiente y que herramientas son utilizadas en distintos países para su control. En el capítulo 2 se describen los estudios similares realizados en años anteriores y además, se incluye la teoría de las técnicas y metodologías

utilizadas en el desarrollo de este trabajo. En el capítulo 3, se describen las fases del experimento, cuya ejecución queda descrita en el capitulo 4, en donde también se analizan los resultados obtenidos. Finalmente, en las conclusiones se realiza una discusión del tema desarrollado.

Capítulo 1

Definición del Problema

1.1. Identificación del Problema

Los incendios forestales se definen según la CONAF como: *"un fuego que [...] se propaga sin control en terrenos rurales, a través de vegetación leñosa, arbustiva o herbácea, viva o muerta."* [1], y son originados en su mayoría por la acción del hombre y en menor parte por rayos de electricidad producidos durante tormentas eléctricas. Su rol en la naturaleza posee dos aristas. Por un lado, los incendios contribuyen a la mantención de la diversidad y salud en los bosques, puesto que eliminan especies que son más susceptibles al fuego e impulsan adaptaciones en aquellas especies que sobreviven; además, cambian la composición de los bosques, creando aperturas capaces de albergar nueva flora y fauna. Por otra parte, los incendios forestales ponen en riesgo el hábitat del ser humano, aumentando la polución del aire, alterando el ciclo del agua y cambiando la composición del suelo. Es por esta razón que al enfrentar este tipo de eventos naturales algunos países tienen el cuidado de alcanzar un balance entre los daños colaterales causados y los beneficios que recibe el bosque en este ciclo natural.

En diversos países, la administración de los recursos forestales está a cargo de una entidad dependiente del estado. En Estados Unidos se encuentra la *U.S Forest Service*; en Canadá cuentan con el *Natural Resource Canada* y el *Canadian Interagency Forest Fire Centre*. las

dos primeras organizaciones realizan labores de mantención y cuidado de la salud de los bosques, mientras que el *Canadian Interagency* se especializa en proveer diversos tipos de servicios para el combate de los incendios forestales. En Chile, la administración de recursos forestales está en manos de la Corporación Nacional Forestal (CONAF), entidad que es responsable de: *"administrar la política forestal de Chile y fomentar el desarrollo del sector"* [2]. Todas estas organizaciones, en sus respectivos países, están a cargo de la prevención y monitoreo de incendios forestales, así como también del registro y almacenamiento de aspectos relevantes ligados a estos tipos de siniestros, como lo son: superficie afectada, fecha y lugar de ocurrencia, entre otros.

Según registros de la CONAF, en Chile ocurren entre 5.000 a 7.000 incendios forestales por temporada, los cuales en promedio afectan una superficie de 52.000 hectáreas anualmente. Posiblemente, unos de los eventos más recordados, y que captó atención a nivel mundial, fueron los múltiples incendios forestales ocurridos en el sector central del Chile durante los meses de enero y febrero del año 2017. Esta serie de incendios fue posteriormente catalogado como una tormenta de fuego la cual consumió aproximadamente 590.000 hectáreas [3]. Afortunadamente, la frecuencia de estos catastróficos acontecimientos es muy baja; sin embargo, estos hechos llaman rápidamente la atención nacional e internacional puesto que el daño causado alcanza grandes proporciones.

Mediante el estudio del comportamiento de incendios forestales, se han desarrollado herramientas computacionales que permiten a las autoridades actuar proactivamente, apoyando la toma de decisiones, la administración y despliegue de recursos antes de que la situación ocurra. En Estados Unidos el *U.S Forest Service* cuenta con un servicio especializado de entrega de información en lo que respecta a: (1) información climática, que permite a expertos determinar los riesgos de un incendio e identificar áreas susceptibles, (2) predicciones respecto al estado de los combustibles y (3) avance geográfico y comportamiento de un incendio.

En el *Canadian Intragency Forest Fire Centre* de Canadá (país considerado líder en la creación de estándares para la mantención sustentable de bosques), se han desarrollado sistemas predictivos cuyo principal objetivo es analizar el comportamiento de incendios y sus potenciales riesgos. El sistema más utilizado en el país es el *Canadian Forest Fire Danger Rating System* (CFFDRS), que provee de métricas para medir el peligro potencial de incendios en

áreas rurales y bosques. Este sistema es apoyado por las siguientes tres herramientas:

- *Forest Fire Weather Index (FWI):* permite medir cambios diarios durante la propagación de un incendio.
- *Forest Fire Behavior Prediction:* estima la tasa de propagación de incendios, el consumo del combustible y la intensidad que puede alcanzar.
- *Fire Effect Model:* permite analizar los efectos inmediatos del incendio y su impacto ecológico en la vegetación.

En Chile, la CONAF ha implementado el Índice de Riesgo de Incendios Forestales, el cual a través de variables meteorológicas y el tipo de vegetación, indica qué tan probable es que se produzca un incendio. El índice tiene 6 categorías de riesgo: nulo, muy bajo, bajo, medio, alto y extremo; a cada uno de estos índices se le asigna un color que luego es utilizado para indicar en un mapa el índice de riesgo asociado a cada zona geográfica del país. Esta representación se crea mediante *Carto*, un software que permite la visualización y análisis de información geográfica.

El fortalecimiento de los sistemas predictivos es una labor que se encuentra en desarrollo en el país. Actualmente CONAF, en conjunto con la Dirección Meteorológica de Chile, se ha planteado la idea de lograr una alianza con Canadá con el objetivo de seguir perfeccionando su labor en el área.

El apoyo que estas herramientas y sistemas entregan resulta fundamental si se tiene en consideración que en el control y combate de incendios forestales se invierte una gran cantidad de recursos económicos. En el caso de Canadá, se han invertido anualmente entre 500 millones y 1 billón de dólares [4] en el combate de incendios. Por otro lado, Estados Unidos prevé que para el año 2025 invertirá aproximadamente 1.8 billones de dólares [5]. En cuanto Chile el año 2016 invirtió un total de 14 mil millones de pesos [6] (21,5 millones de dólares) en esta temática.

Cada herramienta y sistema mencionados anteriormente cuenta con distintos niveles de complejidad en cuanto a su desarrollo e implementación. Las herramientas que modelan la potencial evolución de un incendio requieren de una gran capacidad de cómputo para desplegar sus resultados, debido a que son muchas las variables que inciden en la propagación, situación que dificulta el trabajo con ellas en tiempo real.

En la última década, se han propuesto nuevos modelos predictivos basados en *machine learning*. Esta subárea de la Inteligencia Artificial se encarga de estudiar algoritmos en los que el aprendizaje se lleva a cabo mediante la observación de datos históricos. Los tipos de algoritmos de esta rama se clasifican en las siguientes categorías: (1) Aprendizaje Supervisado, donde el algoritmo crea una relación entre la entrada y la salida deseada; (2) Aprendizaje no Supervisado, en el que se busca organizar el *set* de datos para encontrar estructuras y describirlas, (3) Aprendizaje Semi-Supervisado que combina ambas técnicas anteriores; (4) Aprendizaje por refuerzo, en el que el algoritmo trata de cumplir una tarea guiado por la retroalimentación recibida de su entorno; y (5) Transducción que es similar al aprendizaje supervisado, solo que no crea una relación explícita entre la entrada y la salida del algoritmo.

La predicción de la superficie afectada por un incendio se aborda con algoritmos de Aprendizaje Supervisado, ya que a partir de distintas variables como lo son las condiciones meteorológicas, ubicación geográfica y estado de los combustibles, se realiza una estimación de la superficie afectada por un incendio forestal. La principal ventaja de estos modelos radica en que utilizan datos de fácil acceso y bajo costo.

Considerando los últimos acontecimientos ocurridos en Chile respecto a los incendios forestales, y debido a que en este país el desarrollo de modelos predictivos siguen en etapa de perfeccionamiento, es que en esta memoria se estudiarán los modelos basados en *machine learning* que predicen la superficie quemada por este tipo de incendios. El objetivo de esta investigación es contribuir con mayor información al estudio y aplicación de estos modelos predictivos en incendios forestales ya que la literatura en este ámbito es escasa.

Cabe mencionar que en la elaboración de esta memoria no se pretende desarrollar un sistema que reemplace a los que ya se encuentran implementados en entidades como la *U.S. Forest Service*, *Natural Resources Canada* o CONAF. De lo contrario, se busca proveer con

sustentos técnicos que permitan a estas organizaciones evaluar la factibilidad al implementar dichos modelos en su organización.

En el desarrollo de la memoria, se analizará el comportamiento de una de las mejores técnicas para resolver el problema, las Máquinas de Soporte Vectorial (SVM). Además, se compararán los resultados obtenidos por el modelo propuesto con aquellos registrados en la literatura.

1.2. Objetivo General

Contribuir con un nuevo estudio en lo que corresponde a la predicción de incendios forestales utilizando técnicas de *machine learning*, para proveer de una base que permita determinar la factibilidad de ser implementado.

1.3. Objetivos Específicos

- Modelar y evaluar la precisión alcanzada por una Máquina de Soporte vectorial al predecir la superficie de incendio forestal, para verificar la bondad del modelo.
- Comparar los resultados obtenidos con estudios similares, para determinar si existen mejoras con respecto a éstas.
- Evaluar los cambios en la precisión al plantear la tarea como un problema de clasificación, con el propósito de determinar la viabilidad de esta propuesta.
- Aplicar técnicas de detección de *outliers* para determinar su impacto en los resultados.

1.4. Alcance

- El *set* de datos con el cual se realizarán las pruebas corresponde al del estudio realizado por Cortés y Morais [7], el cual se encuentra disponible para su uso público.

- Los algoritmos a utilizar se desarrollarán en Python, para ello se hará uso de la biblioteca *scikit learn* la cual contiene una implementación las técnicas de *clustering* y *machine learning*.

  Dado a que la memoria constituye una propuesta y ya que no se cuenta con una entidad en la que ésta pueda ser aplicada, la fase de despliegue del modelo de procesos CRISP-DM no se llevará a cabo.

Capítulo 2

Estado del Arte

En este capítulo, se detallan los diferentes estudios que se han realizado en el ámbito de la predicción de incendios forestales. Además, se describen distintas metodologías y técnicas utilizadas en el desarrollo de esta memoria, de modo que el lector pueda tener una mejor perspectiva del trabajo realizado.

2.1. Estudios similares

Cortez y Morais [7] propusieron en el año 2007 un modelo de predicción de incendios forestales basado en el uso de técnicas de minería de datos. Para esto, utilizaron datos de incendios forestales ocurridos en el parque *Montesinho* ubicado en Portugal, de los cuales se tenía registro de: (1) su ubicación geográfica, (2) el día y mes de ocurrencia, (3) cuatro componentes del sistema *Fire Weather Index* (FWI), (4) temperatura ambiental, (5) humedad relativa del ambiente, (6) velocidad del viento, (7) precipitaciones, y (8) el área quemada en hectáreas. Se realizó una fase de pre-procesamiento de datos con el propósito de dejar los datos en un formato aceptado por el modelo, lo cual incluye la transformación de las variables día y mes utilizando la codificación *1-of-C* y la estandarización de las variables. Además, se transformó el área utilizando la función logarítmica $y = \log(x + 1)$, ya que la mayoría de los registros contaba con área 0. Las pruebas se realizaron con cinco técnicas de minería de

datos: máquinas de soporte vectorial, regresión múltiple, árboles de decisión, bosques aleatorios y redes neuronales. Cada una de ellas se entrenó con cuatro subconjuntos del *set* de datos original catalogados por los autores de la siguiente forma: (1) STFWI que incluye la ubicación geográfica, el dia y mes de ocurrencia y las cuatro componentes del FWI, (2) STM compuesto de la ubicación geográfica, el día y mes de ocurrencia y las cuatro variables meteorológicas, (3) FWI que sólo incluye las componentes FWI, y (4) M que utiliza solamente las variables meteorológicas. El mejor resultado fue obtenido por una SVM con *kernel RBF* utilizando sólo variables meteorológicas, cuya precisión fue del 61 % al tolerar un error de 2 hectáreas en la predicción. Cortez y Morais, fueron los primeros en proponer un modelo que utilizara sólo datos meteorológicos (temperatura, humedad relativa, velocidad del viento y precipitaciones) para predecir el tamaño de un incendio forestal, datos a los que se les asocia con un bajo costo de extracción.

En el año 2011 Yang Poh Yu [8], basándose en los resultados de Cortez y Morais, proponen un modelo hibrido de minería de datos. A diferencia del estudio anterior, los autores deciden aplicar técnicas de *clustering* sobre los registros, con el propósito de generar grupos de datos con similares características. A ellos se les asigna una etiqueta la cual indica el tamaño del incendio cualitativamente, transformándose en la nueva variable a predecir por los modelos. A diferencia de [7] los autores deciden trabajar únicamente con las variables meteorológicas, las que son estandarizadas y divididas en conjuntos de entrenamiento y prueba. En la fase de *clustering* los autores entrenan un *Self Organizing Map* (SOM) sobre el *set* de entrenamiento, el cual agrupa a los registros más cercanos según distancia euclidiana. Luego, cada registro del *set* de prueba es asignado a un *cluster* mediante SOM. Una vez obtenidos los *clusters* se asigna a cada registro una de las siguientes cuatro etiquetas: pequeño; mediano; grande y extremadamente grande, según el criterio de la Regla Empírica. Finalmente, se entrena una *Back Propagation Neuronal Network* (BPNN) en cada *cluster* encontrado utilizando los registros del *set* de entrenamiento, y se mide la precisión del clasificador usando el *set* de prueba. Análogamente, se generó un Sistema Basado en Reglas mediante la extracción de reglas IF-THEN en cada *cluster*. Los mejores resultados fueron obtenidos por la BPNN, puesto que el sistema basado en reglas falla al clasificar una muestra que se encuentre fuera del ámbito de las reglas encontradas.

En el año 2014, Guruh Fajar Shidik [10] proponen un método hibrido al igual que Yang Poh Yu. Los autores aplican una fase de *clustering* de datos utilizando *Fuzzy C-means* con el fin de generar nuevas etiquetas que cumplen el mismo propósito del estudio anterior. Previo a la aplicación de este algoritmo el *set* de datos es separado en dos conjuntos, uno que contiene los registros con área 0 y otro con áreas mayores a este valor. Al primer conjunto de datos se le asigna la etiqueta *No burn*, en cambio en el segundo conjunto se generaron dos clusters mediante la aplicación de *Fuzzy C-means*. A estos *clusters* se les asignaron las etiquetas *Light Burn* y *Heavy Burn*. Después de realizada estas modificaciones en el *set* de datos, se efectuó el entrenamiento de una BPNN y una SVM, las que alcanzaron una precisión del 97,5 % y 91,3 % respectivamente. Si bien ambos modelos realizan exitosamente la clasificación de muestras, es importante notar que la asignación de las etiquetas *Light Burn* y *Heavy Burn* parece ser arbitraria, puesto que los autores no realizan un análisis de las características que presentan las agrupaciones obtenidas mediante el algoritmo *Fuzzy C-means*.

El trabajo publicado el 2012 de A. Murat Özbayoğlu [9] se diferencia de los anteriores puesto que utilizan un nuevo *set* de datos, el cual contiene registros de incendios forestales con las siguientes características: (1) humedad relativa, (2) velocidad del viento, (3) temperatura ambiente, (4) *aspect*, (5) *tendency*, (6) estación del año, (7) hora, (8) especie del árbol, (9) cantidad de árboles por unidad de área y (10) área quemada. Al igual que en [8] y [10] se generaron etiquetas para identificar el área quemada mediante el uso de algoritmos *clustering*, debido a que el 80 % de los valores de este atributo son inferiores a una hectárea. En particular se utilizaron dos algoritmos: *K-means* para el set de prueba, y *Fuzzy C-means* para el *set* de entrenamiento y de validación. Los modelos utilizados en este estudio fueron: *Multi Layer Perceptron Network* (MLP); *Radial Basis Function Network*; y SVM. Estos se evaluaron utilizando distintas configuraciones de atributos y valores de los hiper-parametros k y c de las técnicas de *clustering*. Los mejores resultados de este estudio fueron alcanzados por la red neuronal MLP, entrenada con registros que contenían dos atributos (humedad y velocidad del viento), clasificando 65.63 % de los registros exitosamente.

En las figuras 2.1 y 2.2 se comparan las distintas precisiones alcanzadas por cada propuesta revisada, observando que las técnicas híbridas presentan mejores resultados, siendo sólo superadas por las redes neuronales.

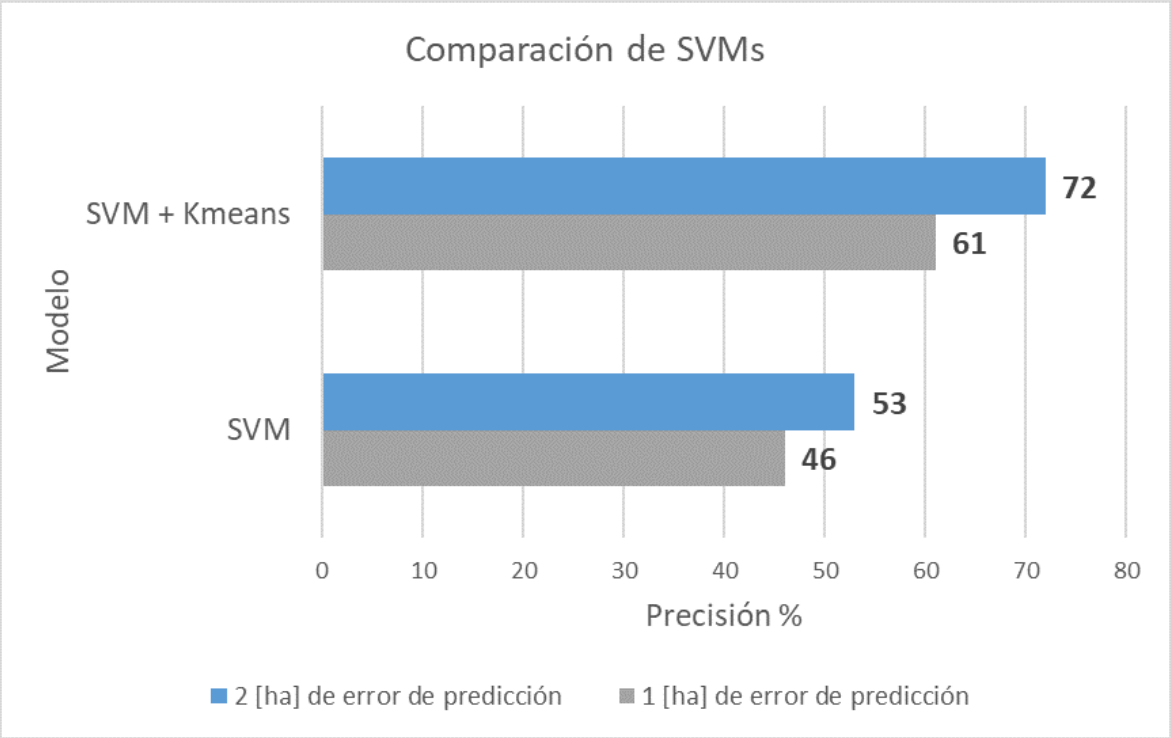


Figura 2.1: *Precisión de distintas SVMs bajo diferentes configuraciones (fabricación propia).*

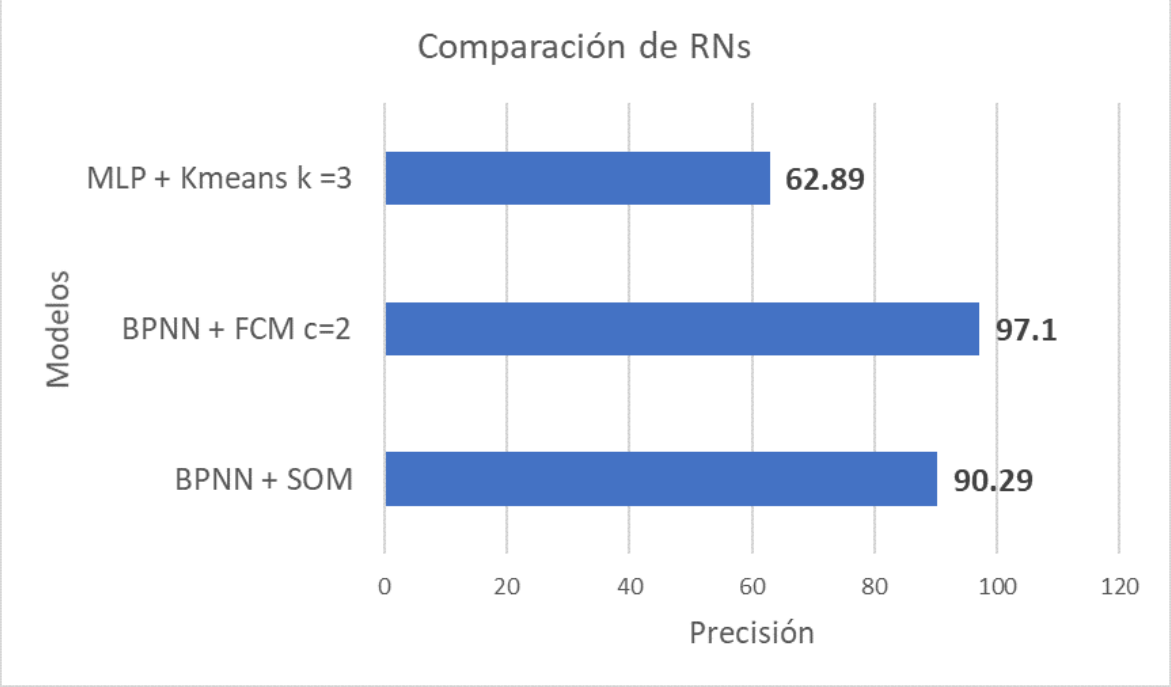


Figura 2.2: *Comparación de la precisión de RNs (fabricación propia).*

2.2. Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (SVM) son *"un clasificador de margen máximo"* [22], utilizado en tareas de aprendizaje supervisado en el ámbito de las máquinas de aprendizaje. El objetivo de este tipo de tareas es estimar un resultado a través de un conjunto de observaciones [23], llamado *set* de entrenamiento. Cada observación se compone de un conjunto de características (por ejemplo: edad, género, altura), así como también del valor de la variable a estimar (por ejemplo: peso en kilogramos). Los algoritmos de aprendizaje supervisado utilizan este *set* de entrenamiento para crear un mapeo entre el conjunto de características y la variable objetivo, el cual se denomina hipótesis o *learner*. De este modo, el *learner* permite la predicción de la variable objetivo para observaciones de las que sólo se conoce el conjunto de características.

Los principales problemas resueltos por los algoritmos de aprendizaje supervisado son los de regresión y clasificación. Se entiende por problemas de clasificación aquellos en el que la variable que se desea estimar es de tipo categórica, es decir, que pertenece a un conjunto discreto de valores; en cambio en problemas de regresión la variable pertenece a un conjunto numérico de valores continuos.

Para explicar el funcionamiento de las SVMs se utilizará un problema de clasificación binario linealmente separable o, en otras palabras, un problema en el cual existen dos clases distribuidas en el espacio de tal manera que éstas pueden ser separadas por un hiper-plano. En la figura 2.3 (a) se ilustra este tipo de problema, donde se observan dos clases (azul y roja), que pueden ser separadas por una recta, sin embargo, la cantidad de rectas que pueden separar ambas clases son infinitas. Una SVM encuentra un hiper-plano separador que maximiza el margen. Éste, corresponde a la distancia entre la frontera de decisión que separa ambas clases y los ejemplos más cercanos a ella, los cuales se denominan vectores de soporte. En la figura 2.3 (b) se ilustra la solución encontrada por una SVM; en ésta, la recta de color negro representa la frontera de decisión encontrada por el algoritmo.

La representación matemática de este clasificador esta dada por la ecuación 2.1, donde $\vec{w}$ corresponde al vector normal del hiperplano separador, b es el intercepto y $\vec{x}$ es el vector de características.

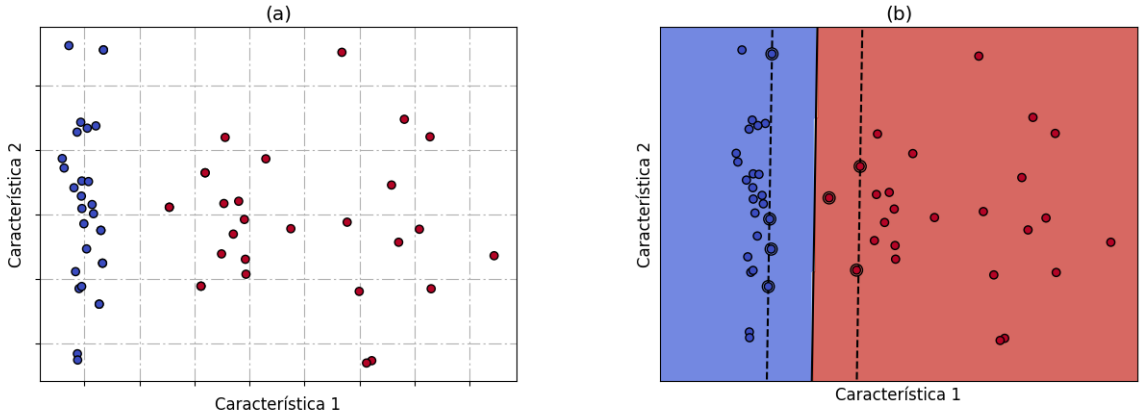


Figura 2.3: *(a) Problema de clasificación linealmente separable, en el cual las clases azul y roja pueden ser separadas por una recta. (b) Solución encontrada por una SVM (fabricación propia).*

$$f(x) = sign(\vec{w}^T \vec{x} + b) \tag{2.1}$$

El clasificador retorna el signo (+1 o -1) que se obtiene al evaluar el vector de características en la ecuación, indicando la pertenencia a una de las dos clases. Siguiendo con el ejemplo de la figura 2.3 las clases roja y azul serían representadas por los valores +1 y -1 o viceversa.

Las SVMs también pueden ser utilizadas para resolver problemas en los que se tienen más de dos clases, conocidos como problemas de clasificación multiclase y, además, para resolver problemas que no son linealmente separables. Los problemas de clasificación multiclase se resuelven utilizando uno de los siguientes métodos: *one-vs-rest* o *one-vs-one*. En *one-vs-rest* se entrena una SVM que determina un hiper-plano separador que distingue entre una de las clases y todo el resto de ellas. En cambio, en *one-vs-one* se entrena una SVM que determina el hiper-plano separador para cada par de clases.

En la figura 2.4 (a) se representa un problema no linealmente separable, en ella se puede observar que las clases azul y roja no pueden ser separadas mediante una recta. Sin embargo, este problema puede transformarse en uno linealmente separable si se proyectan los datos en un espacio de mayor dimensionalidad, tal como se muestra en la figura 2.4 (b).

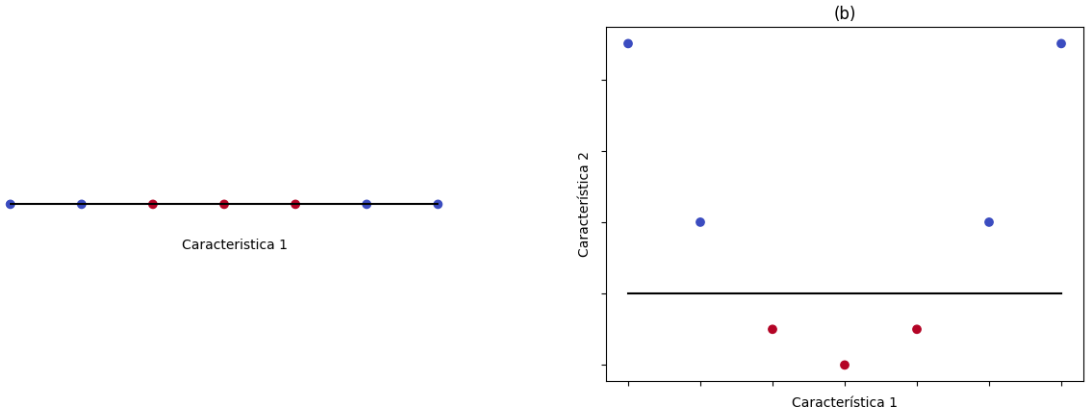


Figura 2.4: *Problema no linealmente separable (a) que al ser proyectado en un espacio de mayor dimensión si lo es (b) (fabricación propia).*

Para realizar esta proyección, se utiliza una función ϕ en la cual se evalúa el vector de características, de este modo la nueva función del clasificador queda definida por la ecuación 2.2.

$$f(x) = sign(\sum_{i} \alpha_i y_i \phi(x_i)^T \phi(x_j) + b) \tag{2.2}$$

Si el producto punto entre $\phi(x_i)^T \phi(x_j)$ puede ser evaluado de forma sencilla y eficiente, entonces no es necesario realizar la transformación del espacio $x \rightarrow \phi(x)$ por el contrario, sólo se debe evaluar este producto punto. Lo anterior, se conoce con el nombre de *Kernel Trick*, el cual consiste en representar el conjunto de características en un espacio de mayor dimensionalidad en el cual los datos pueden ser separados por un hiper-plano.

Un *kernel* es una función que *"corresponde a un producto punto en un algún espacio de características ampliado"*, éste debe ser "*continuo, simétrico, y tener una matriz gram definida positiva"* [22]. En general las dos familias de *Kernels* más utilizados en este ámbito corresponden a las funciones polinomiales y las de base radial.

Por último, las SVM también pueden ser implementadas para resolver problemas de regresión [24]. Para ello, es necesario encontrar los valores de $\vec{w}$ de la ecuación 2.1, asumiendo que los datos pueden ser aproximados mediante una función lineal. Dado que es complejo

que los datos se ajusten a una función de este tipo, los datos son proyectados en un espacio de mayor dimensión, siguiendo la misma lógica que en los problemas de clasificación, donde los productos escalares de la función 2.2 son reemplazados por funciones de *kernel*.

2.3. *Clustering*

Clustering se define como *"el proceso de agrupar un set de objetos físicos o abstractos en clases de objetos similares"* [19], en particular en el campo de la minería de datos este análisis permite descubrir cómo se distribuyen los datos, y además determinar las características de cada clase encontrada [19]. Existen diversos ejemplos de la aplicación de este tipo de análisis, como los son la agrupación de genes con funciones similares, clasificación documentos de la *World Wide Web*, identificación de enfermedades, y la segmentación de grupos de clientes [21].

Los métodos de *clustering* se agrupan en las siguientes categorías [19]: por particionamiento, jerárquicos, basados en densidad, basados en grillas y basado en modelos. Éstas se describen a continuación.

2.3.1. Métodos por particionamiento

Dado un *set* de datos un método por particionamiento clasifica los objetos pertenecientes a este *set* en *k* particiones que representan los *clusters*. Cada uno de éstos debe contener al menos un objeto, y este objeto debe pertenecer a un solo *cluster*, es decir, éstos no se traslapan. La partición inicial es modificada mediante técnicas de relocalización de objetos los cuales mejoran la calidad de la solución encontrada. *K-means* y *k-medoids* son los principales algoritmos de este tipo, los cuales se diferencian en el método de representación de cada *cluster*. *K-means* utiliza el promedio de los objetos pertenecientes a un *cluster* para su representación en cambio, *k-medoids* escoge un objeto cercano al centro del *cluster*.

2.3.2. Métodos jerárquicos

Los métodos jerárquicos son aquéllos que permiten la anidación de *clusters*. Estos métodos se subdividen en dos tipos dependiendo de la forma en que se construyen los clusters; en primer lugar, se encuentran los métodos algomerativos o *bottom-up*, en los que inicialmente cada objeto forma un *cluster*, los cuales se van uniendo hasta formar uno solo que contiene todo el *set* de datos. En cambio, en los métodos divisivos o *top-down* ocurre lo contrario, en la fase inicial todos los objetos forman un solo *cluster*, que luego se va descomponiendo en conjuntos de menor tamaño. El método finaliza cuando cada objeto forma su propio *cluster*.

La principal desventaja de estos métodos es que no permiten la relocalización de objetos, debido a que la unión o separación de los *clusters*, no es reversible. Sin embargo, esta característica le permite disminuir los tiempos de cómputo.

2.3.3. Métodos basados en densidad

Los métodos de particionamiento tienden a construir *clusters* de forma esférica, debido a que utilizan nociones de distancia para evaluar la pertenencia de un objeto a un *cluster*. A causa de esto, es que se desarrollan los métodos basados en densidad, los cuales permiten encontrar *clusters* de formas arbitrarias. Esto se logra contando la cantidad de objetos que se encuentran dentro de un radio definido o vecindad. Si ésta supera un límite definido entonces el *cluster* se expande hasta que esta condición no se cumple.

2.3.4. Métodos basados en grillas

Este tipo de métodos divide el espacio generado por los objetos del *set* de datos en grillas. Luego, es en estas grillas donde se realizan las operaciones que permiten definir los *clusters* del *set* de datos. Este acercamiento repercute directamente en el tiempo de procesamiento, ya que éste depende solamente del número de grillas en que se haya dividido el espacio y no de la cantidad de objetos como en los métodos anteriores.

2.3.5. Métodos basados en modelos

Estos métodos requieren la formulación de un modelo para encontrar cada *cluster*, lo que se puede lograr definiendo una función de densidad que represente la distribución de los objetos en el espacio. Luego, los algoritmos construyen los *clusters* buscando el mejor ajuste de los objetos para el modelo correspondiente.

2.4. Modelos de Proceso Proyectos de Minería de Datos

La minería de datos es un proceso mediante el cual se predice un resultado o comportamiento a través de la identificación de patrones en un conjunto de datos [11]. En los últimos años este proceso ha cobrado importancia dentro de las empresas y organizaciones, puesto que les ha permitido encontrar una nueva fuente de información que apoya la toma de decisiones y además, descubrir un nuevo uso a la gran cantidad de datos que éstas generan día a día. Es por esta razón, que se han propuesto diversos modelos que describen la forma en que se debe ejecutar un proceso de minería de datos para conseguir los resultados deseados. Los modelos que se revisarán a continuación serán CRISP-DM, SEMMA Y KDD.

2.4.1. CRISP-DM

El *Cross-Industry Standard Process for Data Mining* (CRISP-DM) es un modelo de procesos jerárquico, creado el año 1999 por las empresas Daimler Chrysler, SPSS y NCR con el objetivo de proveer un *framework* y guías para realizar proyectos de minería de datos. Este modelo está compuesto de cuatro niveles de abstracción [12], los cuales son: (1) fase, (2) tarea genérica, (3) tarea especializada e (4) instancia de proceso, que estructuran el proceso de minería de datos. Específicamente, este proceso se compone de varias fases, las cuales a su vez están compuestas de varias tareas genéricas quienes describen las acciones a seguir en cualquier tipo de proceso de minería de datos. Del mismo modo, las tareas genéricas se componen de tareas especializadas, que describen cómo se deben ejecutar estas acciones en las tareas generales. Finalmente, en la instancia de proceso se registran las acciones,

decisiones y resultados obtenidos en el proceso de minería de datos.

A continuación, se describirán en grandes rasgos las 6 fases que conforman el proceso CRISP-DM [12]:

1. **Comprensión del Negocio:** una de las fases más importantes, un buen entendimiento del problema y su repercusión para el negocio permite realizar una buena elección de datos y una correcta interpretación de los resultados. En esta fase, se definen los objetivos del proyecto de minería de datos y del negocio, y se describe su funcionamiento antes del inicio del proyecto.

2. **Comprensión de los Datos:** en ésta se construye el set de datos a utilizar. Esta fase involucra las etapas de descripción, exploración y verificación de calidad de los datos

3. **Preparación de los Datos:** fase en la que se prepara el *set* de datos para su uso con las distintas técnicas de minería de datos. Tareas como lo son la selección de atributos, limpieza de datos y formateo de los datos son realizadas.

4. **Modelado:** en esta fase se escogen los modelos a utilizar para resolver el problema definido en la comprensión del negocio. Se ejecutan actividades como la generación de un plan de prueba, la construcción y evaluación del modelo.

5. **Evaluación:** se verifica el cumplimiento de los objetivos establecidos para cada uno de los modelos obtenidos.

6. **Despliegue:** última fase en la que el conocimiento generado por el modelo se integra a los procesos del negocio

2.4.2. SEMMA

SEMMA es un proceso de minería de datos desarrollado por la compañía SAS para el software SAS Enterprise Miner. Su nombre proviene de las cinco fases que conforman el proceso, las cuales se detallan a continuación [13]:

1. **Muestra:** se divide el set de datos original en sub-sets cuyo tamaño permita representar bien la información y, además, un rápido procesamiento.

2. **Exploración:** se centra en descubrir posibles relaciones y anormalidades en el set de datos, de manera que se genere un conocimiento del problema.

3. **Modificación:** se crean, escogen y transforman los datos originales para facilitar la elección de los modelos en la siguiente fase.

4. **Modelado:** se modelan los datos utilizando modelos estadísticos o de máquinas de aprendizaje para predecir el resultado deseado.

5. **Evaluación:** se evalúa la confianza y uso de los resultados obtenidos del proceso de minería de datos.

2.4.3. KDD

Knowledge Discovery in Databases (KDD), como su nombre lo indica, corresponde al proceso de encontrar el conocimiento que yace en un conjunto de datos, los cuales se encuentran almacenados en una base de datos de gran tamaño [14]. Este conocimiento es identificado al encontrar patrones en el conjunto de datos que son inesperados o novedosos y que, además, son útiles para el negocio. Este proceso consta de los siguientes pasos [15]:

1. Identificar y entender los objetivos del usuario final, el dominio de la aplicación y el conocimiento existente antes de la ejecución del proceso.

2. Seleccionar el *set* de datos a utilizar.

3. Preprocesar los datos, eliminando inconsistencias y datos ruidosos; esto, se logra a través de la eliminación de *outliers* y la implementación de estrategias para administrar datos nulos, entre otros.

4. Transformar, de ser necesario, el *set* de datos con el objetivo de facilitar la implementación de los algoritmos, técnicas como la reducción de dimensionalidad y la transformación del tipo de datos son aplicados en este paso.

5. Escoger un método de minería de datos que sea apropiado para cumplir los objetivos identificados.

6. Escoger e Implementar un algoritmo de minería de datos ajustando sus parámetros para cumplir con los objetivos establecidos.

7. Interpretar, evaluar y documentar los patrones encontrados, de manera que se pueda emplear el conocimiento encontrado en el propósito identificado al comienzo del proceso.

2.4.4. Similitudes y diferencias entre los procesos

Los estudios realizados por [16], [17] y [18] indican que estos tres modelos presentan fases que son equivalentes entre sí, las que se encuentran resumida en la figura 2.5. Estos estudios concluyen que KDD presenta menos nivel de detalle que SEMMA y CRISP-DM, siendo este último el que cuenta con más nivel de detalle al describir las tareas que se deben realizar en cada fase del proceso. Si bien SEMMA y CRISP-DM son vistos como una implementación del proceso KDD, los estudios concluyen que CRISP-DM es el modelo de más completo.

Modelos de Procesos de Minería de Datos	KDD	CRISP-DM	SEMMA
Número de pasos	9	6	5
Nombre de cada paso	Desarrollo y comprensión de la aplicación	Comprensión del negocio	-
	Crear un *set* de datos	Comprensión de los datos	Muestra
	Limpieza y preprocesamiento de datos		Exploración
	Transformación de datos	Preparación de los datos	Modificación
	Escoger el cometido de la minería de datos	Modelado	Modelo
	Escoger el algoritmo de minería de datos apropiado		
	Desplegar el algoritmo de minería de datos		
	Interpretar los patrones encontrados	Evaluación	Balance
	Usar el conocimiento encontrado	Despliegue	-

Figura 2.5: *Tabla que refleja las fases equivalentes entre cada proceso de Minería de Datos [18].*

Capítulo 3

Propuesta

En este capitulo se detalla el proceso de minería de datos con el cual se desarrolla el modelo que permitirá realizar la predicción del tamaño de incendios forestales. Además, se especifican las características del *set* de datos utilizado, los *tests* a realizar, su método de evaluación, y las herramientas usadas para su implementación.

3.1. Modelo de proceso de Minería de Datos a utilizar

En la sección 2.4 se detallaron tres de los modelos más utilizados en minería de datos, los cuales son: CRISP-DM, KDD y SEMMA. Para el desarrollo de esta memoria se ha decidido utilizar el modelo de procesos CRISP-DM, dado su mayor nivel de completitud respecto al resto de los modelos. En consecuencia, la propuesta se estructurará en las siguientes fases: (1) Comprensión del Negocio, (2) Comprensión de los Datos, (3) Preparación de los Datos, (4) Modelado, (5) y Evaluación.

3.1.1. Compresión del Negocio

Valoración de la situación actual

En el capítulo 1 se detallaron los aspectos más relevantes del combate y detección de incendios forestales en países como Chile, Canadá y Estados Unidos. Destacando los avances que han hecho las entidades que se encuentran a cargo del monitoreo y combate de incendios forestales en sus respectivos países, especificando el costo monetario que conlleva anualmente enfrentar este tipo de incendios. Finalmente en el caso particular de Chile se destaca el trabajo que se está efectuando en cuanto a la mejora de sus modelos predictivos a través de la generación de alianzas con países como el de Canadá.

Objetivos del Negocio y criterios de éxito

El objetivo principal del proceso de minería de datos consiste en determinar el área quemada por un incendio forestal. Para cumplir con dicho objetivo, se ha definido como criterio de éxito la obtención de modelos cuyas métricas de evaluación sean mejores que las obtenidas en los estudios expuestos en la sección 2.1, los cuales son la precisión para los modelos de clasificación, y el error medio absoluto (MAE) y la raíz cuadrada del error cuadrático medio (RMSE) para el modelo de regresión.

3.1.2. Comprensión de los datos

En esta fase, se realiza la tarea de descripción de datos a través del perfilado de cada una de las características y el posterior análisis de histogramas. Finalmente, se analizan la calidad de los registros para detectar aquellas mediciones que se encuentren en rangos extremos. Esto se logra empleando la detección de *outliers* vía dendrogramas, utilizando como métricas de distancia llamadas euclideana, coseno y Manhattan.

3.1.3. Transformación de los datos

En esta fase se realiza la codificación de las características mes y día a valores discretos, para que puedan ser procesados correctamente por el modelo. Además, se estandarizan los datos, debido a que las características del *set* de datos se encuentran en distintas escalas. Finalmente se seleccionan los parámetros en base a rankings, los cuales se obtendrán utilizando las métricas de Información Mutua y los valores *F-score* del *test* ANOVA, y el valor del test χ^2 las que en general miden las dependencias entre dos variables.

3.1.4. Modelado

Tests a realizar

Los tres *tests* a realizar corresponden a: (1) *Test* de regresión, (2) *Test* de clasificación I y (3) *Test* de clasificación II. El objetivo del *Test* de Regresión es realizar una predicción cuantitativa del área quemada por los incendios. En cambio, los *Tests* de clasificación I y II tienen por objetivo realizar una predicción cualitativa de ella. Para lograrlo, se agruparán los datos aplicando técnicas de *clustering* (para el caso del *Test* de clasificación I) a los que se les asignará una etiqueta, la que posteriormente será predecida por el modelo. El Test de clasificación II se diferencia del anterior en la fase de agrupación de datos, ya que ésta se lleva a cabo agrupando el área por cuartiles. Las técnicas escogidas para realizar *clustering* son: *K-means*, *DBSCAN*, Aglomerativo con métricas *Ward* y *complete*, y *clustering* espectral. Estos tres *tests* se implementan con el propósito de evaluar las ventajas y desventajas que implica utilizar un método por sobre el otro, lo cual se determinará a través de métricas como la precisión del modelo y los tiempos de entrenamiento asociados a cada uno. Por último, pero no menos importante, se compararán los resultados obtenidos con los estudios expuestos en la sección 2.1. El modelo a implementar en los tres *tests* será el de las máquinas de soporte vectorial (SVM) ya que, tal como se revisó en el estado del arte, corresponde a una de las técnicas con mayor precisión, siendo ésta superada únicamente por las redes neuronales. Está última no fue considerada dentro del desarrollo de la memoria, ya que la precisión alcanzada en estudios similares deja un bajo margen de mejora en comparación con los modelos que

utilizan las máquinas de soporte vectorial.

Método de entrenamiento

El entrenamiento de los modelos se realizará con el 70 % de los registros del *set* de datos mientras que el 30 % restante será utilizado como *set* de *testing*. Por otra parte, los hiperparámetros de la SVM serán sintonizados utilizando *K-fold Cross Validation* con 10 folds.

3.1.5. Evaluación

Las métricas utilizadas en la evaluación dependerán del tipo de *test* efectuado. Para el *test* de regresión se utilizará el error medio absoluto y la precisión alcanzada al tolerar un error de predicción de 2 hectáreas, para así poder realizar una comparación con el estudio [7]. Por otra parte para la evaluación de los *test* de clasificación se utilizarán la precisión, *recall* y tiempos den entrenamiento.

3.2. *Set* de datos

El *set* de datos que se utilizará corresponde al usado en [7] el cual se encuentra disponible públicamente en [26]. Éste se compone de 517 registros de incendios forestales ocurridos en el Parque Nacional *Montesinho* ubicado en Portugal. Cada uno de estos registros cuenta con 13 características, enumeradas a continuación: (1) Coordenada X e (2) Y de la ubicación del incendio en una grilla de 9×9 del mapa del parque, (3) mes y (4) día de ocurrencia, (5) temperatura, (6) humedad relativa, (7) velocidad del viento, (8) precipitaciones, (9) *Fine Fuel Moisture Code* (FFMC), (10) *Duff Moisture Code* (DMC), (11) *Drougth Code* (DC), (12) *Initial Spread Index* (ISI) y (13) área quemada. De ellas, X e Y son variables discretas, mes y día son categóricas, y el resto de ellas continuas.

3.3. Hardware y Herramientas a utilizar

Para la ejecución de los modelos se utilizará un equipo con las siguientes características:

- RAM: 8GB
- CPU: Intel(R) Core (TM) i3 1.70[GHz]

Para implementar los modelos se utilizó Python en su versión 2.7 y la biblioteca *scikit-learn* en su versión 0.19.1, la cual contiene los módulos necesarios para realizar la transformación de los datos, junto con implementaciones de SVMs para realizar tareas de regresión y clasificación.

Capítulo 4

Implementación

En este capítulo se realizan las tareas descritas desde la segunda fase de CRISP-DM en adelante, comenzando con la descripción de datos y su transformación, para finalmente elaborar el modelo presentando sus resultados y conclusiones.

4.1. Comprensión de los datos

Para el desarrollo del experimento se utilizó el *set* de datos de la publicación [7], el cual corresponde a registros de incendios ocurridos en el Parque Nacional *Montesinho* ubicado en Portugal durante los años 2000 y 2003. Este *set* de datos cuenta con un total de 517 registros, cada uno con 13 características descritas a continuación:

- **X:** coordenada X de la ubicación del incendio en el mapa del parque Monstesinho.
- **Y:** coordenada Y de la ubicación del incendio en el mapa del parque Montesinho.
- **Mes:** mes en que ocurrió el incendio.
- **Día:** día de la semana en que ocurrió el incendio.
- **Temperatura:** temperatura del ambiente, medida en °C.

- **Humedad Relativa:** humedad relativa del ambiente, medida en porcentaje.
- **Velocidad del Viento:** velocidad del viento, medida en [km/h]
- **Precipitaciones:** nivel de precipitaciones, medida en [mm/m2]
- ***Fine Fuel Mosture Code* (FFMC):** humedad de la capa de suelo ubicada entre la superficie hasta 1.2 cm de profundidad.
- ***Duff Moisture Code* (DMC):** humedad de la capa de suelo ubicada entre 1.2[cm] hasta 7[cm] de profundidad.
- ***Drought Code* (DC):** humedad de la capa de suelo ubicada a partir de los 7[cm] de profundidad.
- ***Initial Spread Index* (ISI):** tasa con la cual se propaga el incendio.
- **Área:** área total quemada por el incendio, medida en hectáreas [ha].

Los autores de [7] clasifican estas características en categorías, el primer grupo corresponde a las características (1) espacio-temporales, conformado por los atributos X, Y, mes y día. Luego se encuentran las características (2) meteorológicas que incluyen temperatura, humedad, velocidad del viento y las precipitaciones; y finalmente se encuentra las características del sistema (3) *Fire Weather Index* conformada por los códigos FFMC, DMC, DC e ISI. Esta clasificación será utilizada para poder organizar de mejor forma la descripción de los datos.

4.1.1. Características Espacio – Temporales

Las coordenadas X e Y, que indican la ubicación del incendio dentro del parque, son extraídas del mapa de la figura 4.1, en el cual se realiza la división del terreno en una grilla de 9×9, lo cual permite identificar las zonas a través de la tupla (X,Y). En la figura 4.2 se indican las zonas en las que ocurrieron más incendios durante los años 2000 y 2003. En ésta se observa que en las tres zonas con más incendios, más de la mitad de éstos resultaron ser de un tamaño menor a 100[m^2].

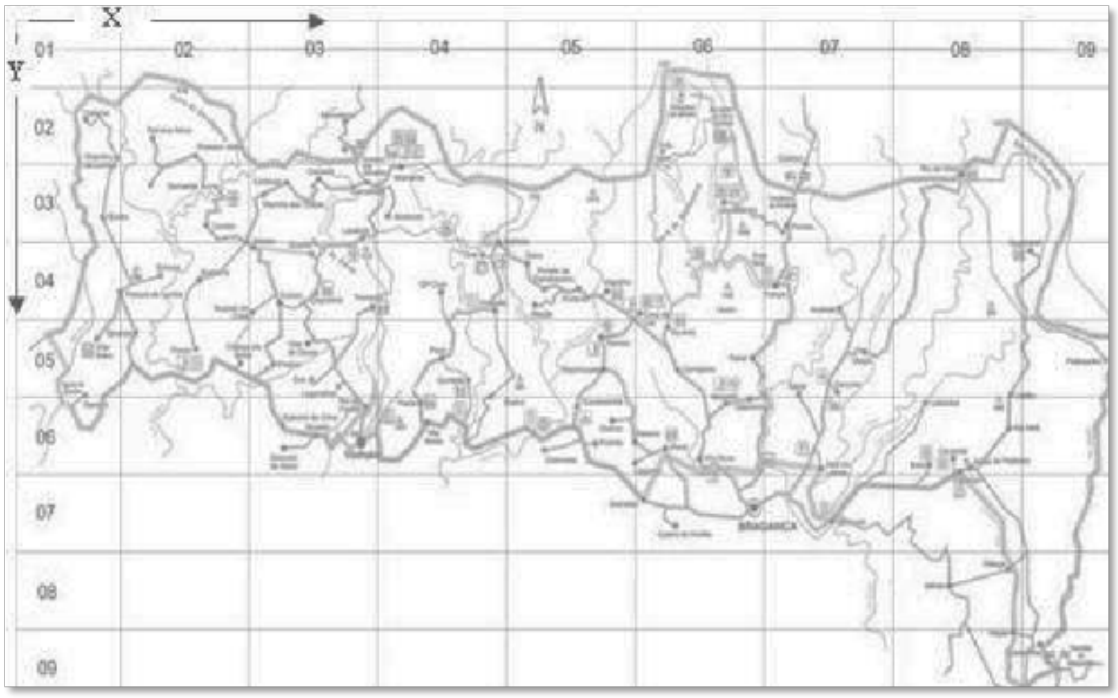


Figura 4.1: *División en grillas del terreno del Parque Natural Montesinho [7].*

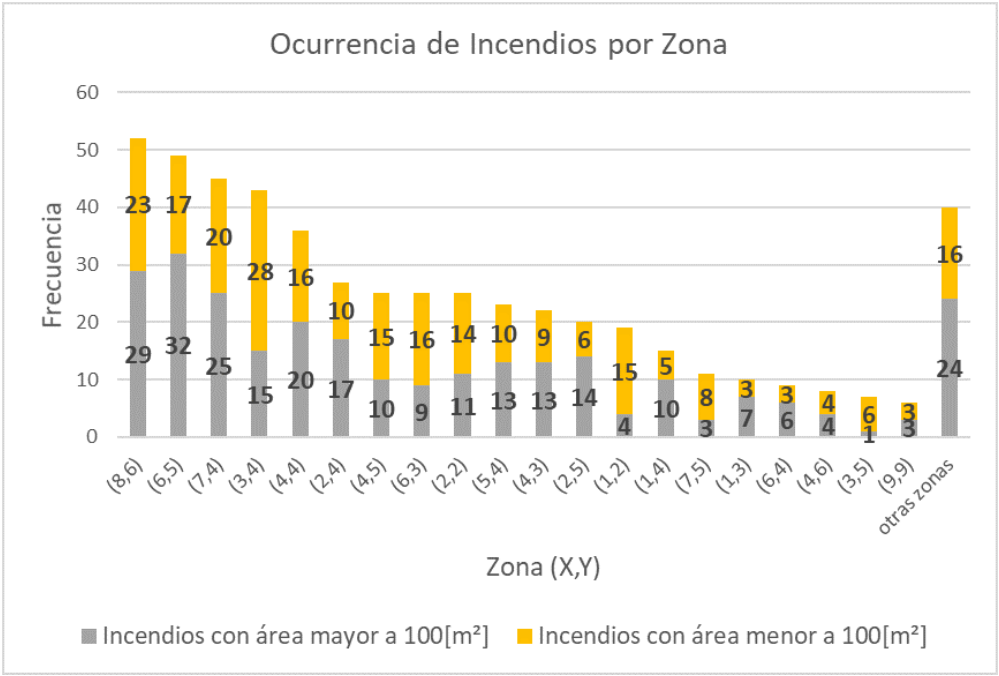


Figura 4.2: *Distribucíon de los incendios por cada zona del mapa del Parque Natural Montesinho (fabricación propia).*

Las características mes y día corresponden a variables categóricas que entregan una noción temporal de la ocurrencia de los incendios. En figuras 4.3 y 4.4 se visualiza la ocurrencia de incendios por día y mes, respectivamente. En éstas figuras se observa que los incendios ocurren con más frecuencia entre los días viernes, sábado y domingo, los cuales coinciden con días en que la visita de público incrementa, por ser días de fin de semana. Esta consideración es importante, debido a que la mayor parte de los incendios forestales son originados por la acción del hombre.

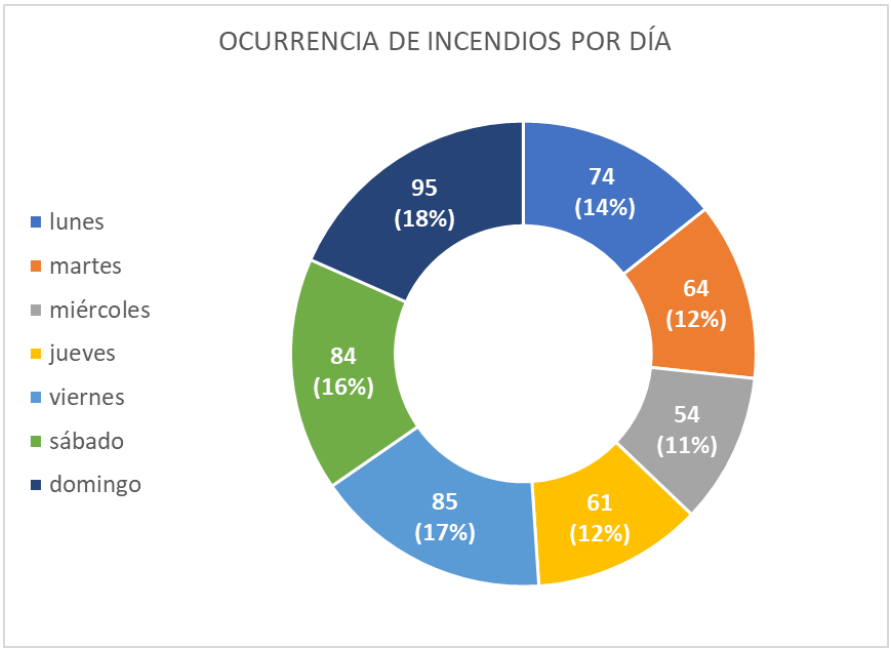


Figura 4.3: *Distribución de los incendios por cada día de la semana (fabricación propia).*

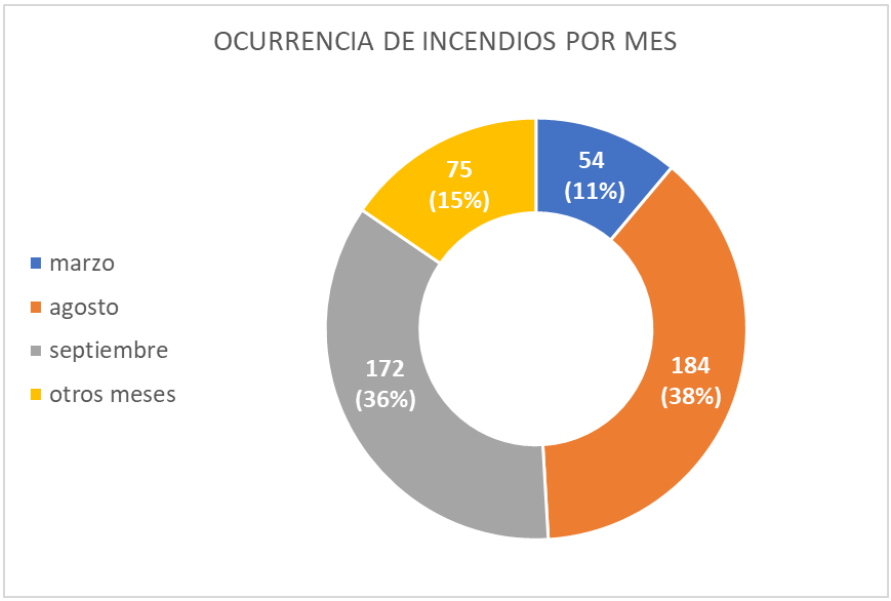


Figura 4.4: *Ocurrencia de incendios por mes (fabricación propia).*

Por otra parte, al analizar los meses se observa que los incendios se concentran mayoritariamente en las estaciones de verano e inicios de otoño del hemisferio norte, estaciones en las que se registran las temperaturas más altas en la región [25]. Los meses que no figuran en el gráfico se agruparon bajo la etiqueta “otros meses” debido a que la ocurrencia de incendios en aquellos periodos era baja con respecto a los meses de mayor ocurrencia, representando sólo 15 % del total de incendios.

4.1.2. Características Meteorológicas

Este conjunto de características entrega información respecto a las condiciones en que se encuentra el ambiente al momento de ocurrir un incendio. Los datos fueron recolectados mediante las mediciones realizadas por la Bragaça Politechnic Institute utilizando la estación meteorológica que se ubica en dependencias del parque. El cuadro 4.1 contiene el perfilado de estas características en el cual se detallan los valores mínimos y máximos, medianas, promedios y cantidad de valores únicos. Además, en la figura 4.5 se pueden observar los histogramas de cada una de estas variables.

	Temperatura	Humedad Relativa	Velocidad del Viento	Precipitaciones
Mínimo	2.2	15	0.4	0
Máximo	33.3	100	9.4	6.4
Promedio	18.88	44.28	4.01	0.02
Mediana	19.3	42	4	0
Valores Únicos	192	75	21	7

Cuadro 4.1: *Perfilado de las características meteorológicas (fabricación propia).*

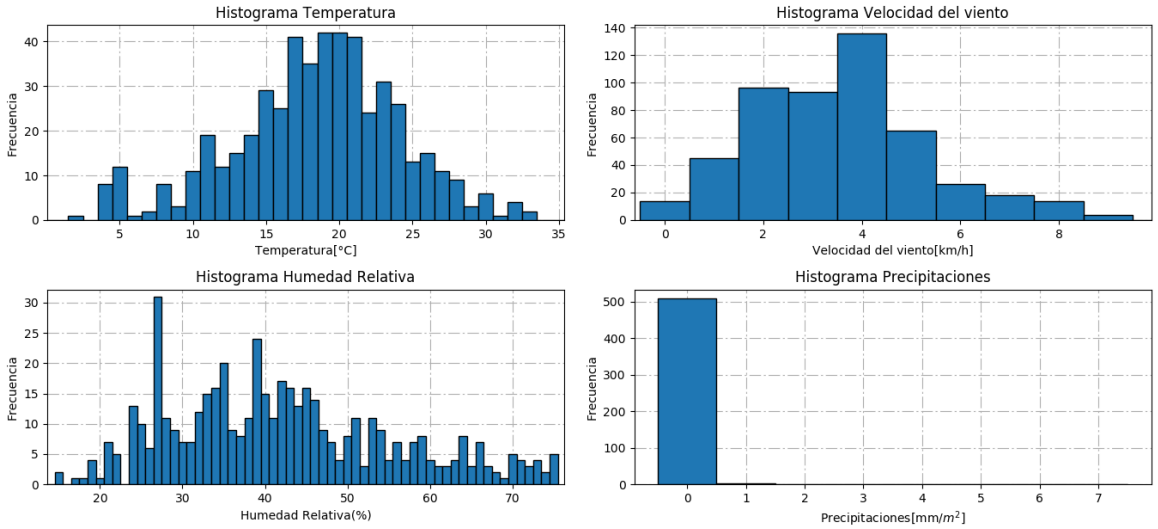


Figura 4.5: *Histogramas de las características meteorológicas, donde se observa un sesgo en las precipitaciones (fabricación propia).*

El histograma de las precipitaciones indica que esta característica presenta un sesgo en torno a 0, ya que como lo indica su mediana más de la mitad de los registros de incendios forestales se encuentran con un valor nulo de precipitaciones. Este hecho podría tener relación con que el 74 % de los incendios se originaran en verano, estación que se caracteriza por sus bajas precipitaciones y altas temperaturas. El resto de las características meteorológicas no presentan un sesgo tan notorio como el observado en las precipitaciones, puesto que no se registran mediciones con valores extremos y de baja frecuencia.

4.1.3. Característias pertenecientes al *Fire Weather Index* (FWI)

El *Fire Weather Index* (FWI) corresponde a un *"sistema que consiste de seis componentes que miden los efectos de la humedad de los combustibles y el viento en el comportamiento de un incendio"*. Las componentes registradas en el *set* de datos incluyen tres códigos que miden la humedad de combustibles en distintos niveles de profundidad del suelo, éstos son: *Fine Fuel Moisture Code* (FFMC); *Duff Moisture Code* (DMC) y *Droguth Code* (DC). Valores bajos de los codigos FFMC, DMC y DC indican que los niveles de humedad de los combustibles del bosque son altos, mientras que valores altos indican bajos niveles de humedad. El cuarto componente corresponde al *Initial Spread Index* (ISI), índice que representa la tasa esperada de propagación del incendio. En la figura 4.6 se muestran los datos necesarios para obtener cada métrica del sistema; en ésta se observa que las cuatro componentes registradas en el *set* de datos son derivadas de las características meteorológicas.

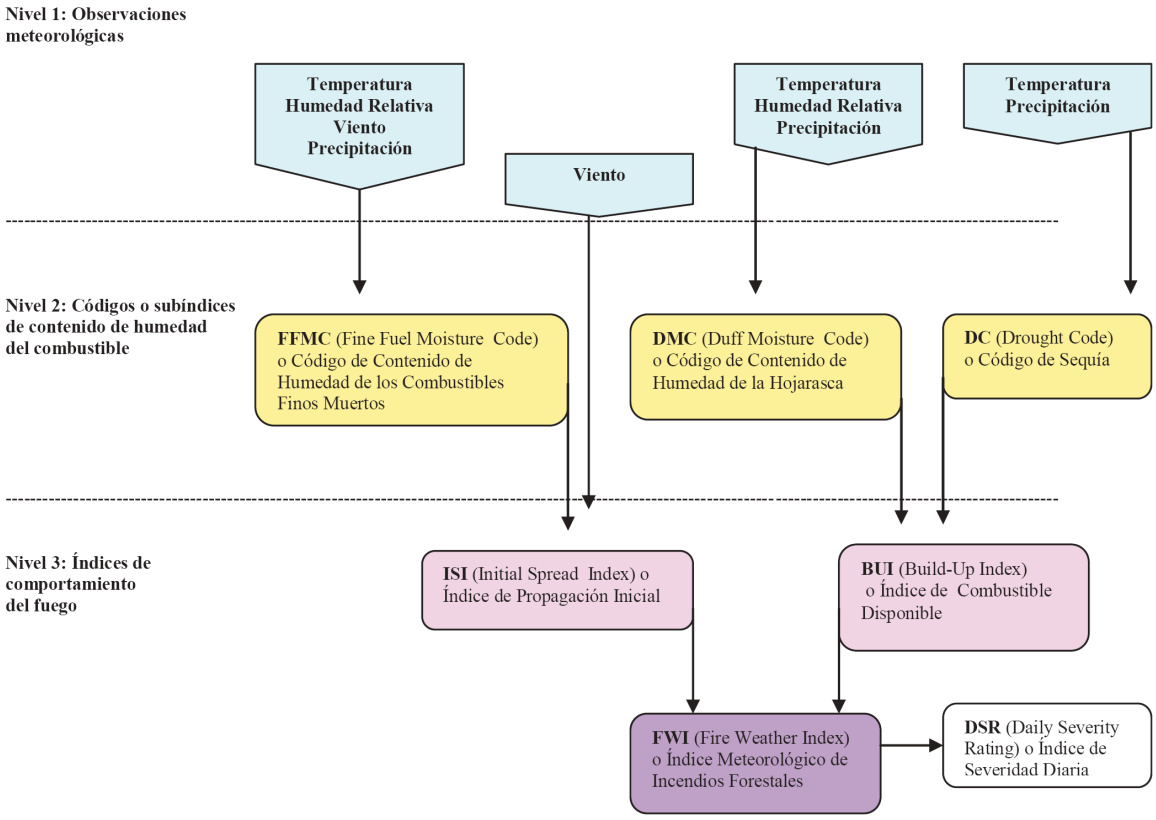


Figura 4.6: *Esquema para obtener cada componente del sistema FWI* **??**.

Al igual que con las características meteorológicas, se ha realizado un perfilado y construido los histogramas de estos datos, los cuales se encuentran en el cuadro 4.2 y la figura 4.7, respectivamente.

	FFMC	**DMC**	**DC**	**ISI**
Mínimo	18.7	1.1	7.9	0
Máximo	96.2	291.3	860.6	56.1
Promedio	90.64	110.87	547.94	9.02
Mediana	91.6	108.3	664.2	8.4
Valores Únicos	106	215	219	119

Cuadro 4.2: *Perfilado de las características del sistema FWI (fabricación propia).*

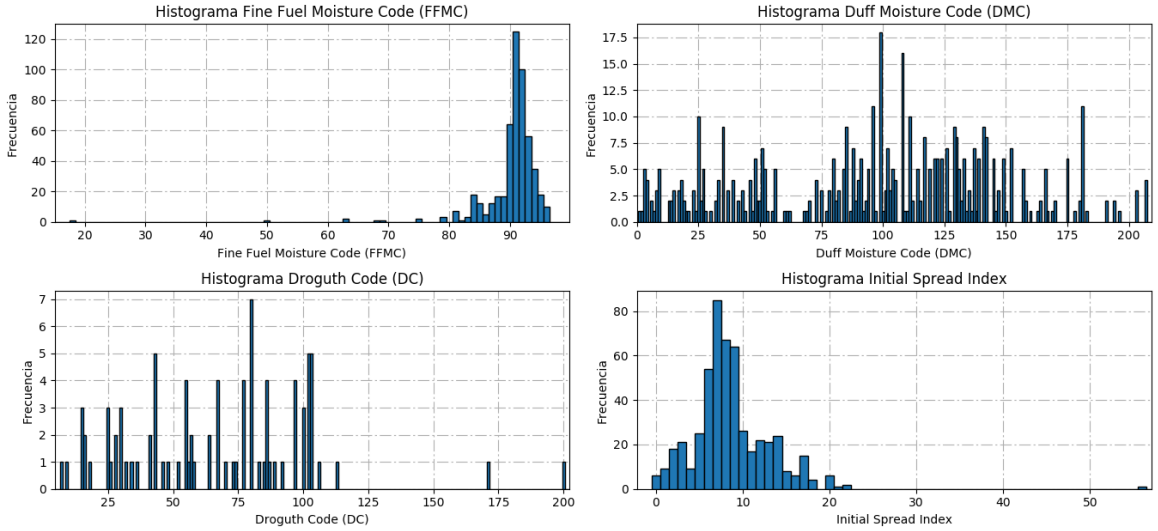


Figura 4.7: *Perfilado de las características del sistema FWI (elaboración propia).*

El valor máximo que puede alcanzar la componente FFMC según su formula matemática es 101, mientras que las ecuaciones para calcular el resto de las variables del sistema no presentan un valor máximo. Sin embargo existen un límite a partir del cual cada código comienza a representar un nivel alto de riesgo, los cuales son: código FFMC 85, DMC 28, DC 190 e ISI 5 [28]. En el cuadro 4.2 se observa que más de la mitad de los registros

presentan niveles de riesgo altos dentro de la escala nombrada anteriormente, los cuales llegan a ser extremos en el caso de las componenetes DMC y DC.

En los histogramas se puede observar que los valores mínimos y máximos de características como FFMC, DC e ISI son casos particulares que ocurren con una baja frecuencia. En el caso del FFMC y ISI, estos se asocian a incendios cuya área quemada no supera los 100[m^2]. Lo que llama particularmente la atención en el caso de ISI puesto que mayores valores indican que el incendio abarca una mayor área en un lapso de tiempo menor, lo cual puede sugerir que hubo una acción de supresión y control del incendio eficaz, o que las demás condiciones no fueron propicias para su expansión.

4.1.4. Área

Como su nombre lo indica corresponde al área afectada por el incendio medida en hectáreas. En este *set* de datos en particular, aproximadamente la mitad de los incendios forestales registrados con un tamaño menor a 100[m^2] fueron codificados con el valor 0, lo cual origina la situación observada en la figura 4.8, en donde la mayor cantidad de incendios es de un tamaño menor de 100[m^2].

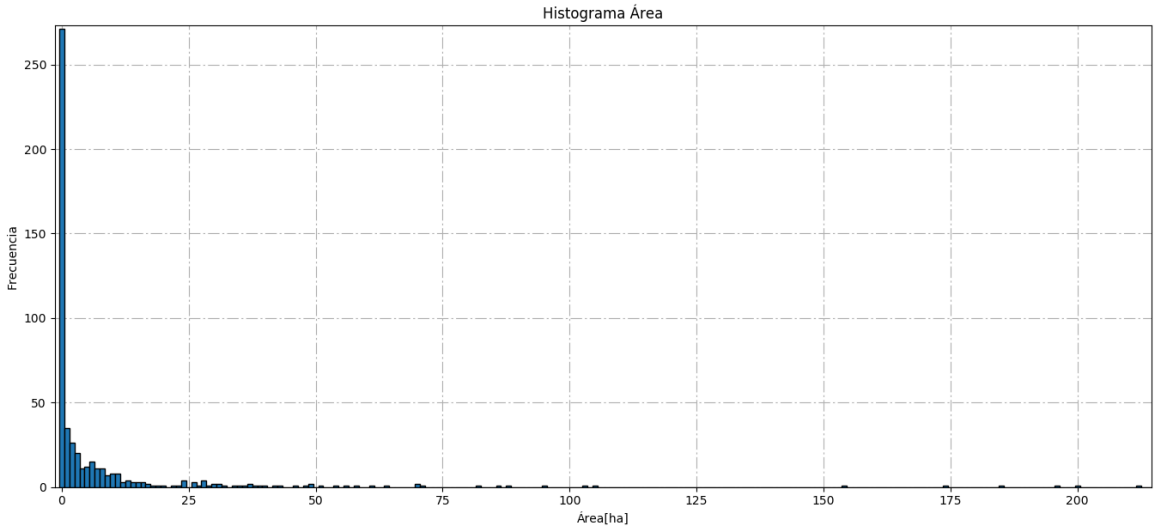


Figura 4.8: *Histograma del área quemada por los incendios forestales del Parque Natural Monstesinho (fabricación propia).*

El sesgo que presenta esta variable, corresponde a una característica bastante común a nivel internacional, en donde los incendios que abarcan grandes zonas como lo ocurrido en Chile entre enero y febrero del 2017 tienen una muy baja frecuencia. Es por ello, que los valores cercanos al máximo de 1090.84 hectáreas, detallados en el cuadro 4.3 pueden ser considerados como valores *outliers*. Con el objetivo de lograr una mejor comprensión de la distribución de los valores del área, la figura 4.9 muestra la proporción específica del tamaño de los incendios en el *set* de datos. De un total de 571 siniestros registrados, un 48 % de ellos abarcan un área inferior a los 100[m^2] y el 52 % de éstos cubre un área mayor que ésta.

	Área
Mínimo	0
Máximo	1090.84
Promedio	12.84
Mediana	0.52
Valores Únicos	251

Cuadro 4.3: *Perfilado del área (fabricación propia).*

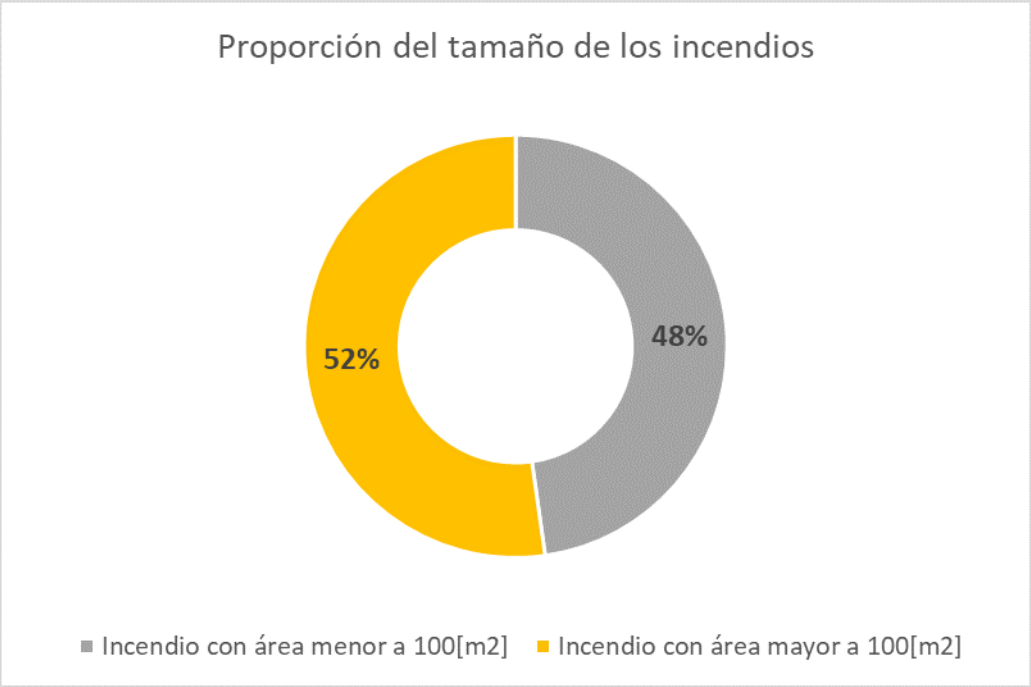


Figura 4.9: *Proporción del tamaño de los incendios dentro del Parque Nacional Montesinho (fabricación propia).*

4.2. Transformación de datos

4.2.1. Transformación de las características mes y día

La transformación de las características ordinales mes y día a variables discretas, se realiza con el objetivo de representar estas dos columnas de datos en un formato procesado por una SVM. Dos de los métodos mas comunes para realizar esta codificación son: *Label Encoder* y *One Hot Encoder*. La técnica de *Label Encoder* consiste en asignar a cada una de las N categorías un valor entre 0 y $N - 1$. Por otra parte, *One Hot Encoder* crea una nueva variable binaria por cada uno de las N categorías, indicando la pertenencia a una categoría mediante el valor 1, y 0 en el caso contrario. Se ha decidido utilizar *Label Encoder* ya que en conjunto, se tienen 19 etiquetas para las variables mes y día; y el utilizar *One Hot Encoder* incluiría 19 nuevas características al *set* de datos, aumentando su dimensionalidad de 13 a 32. Lo

que podría conllevar a un mal rendimiento del modelo predictivo y algoritmos de *clustering*, puesto que los datos podrían tener una dispersión mayor en este nuevo espacio dimensional. De este modo, los meses son enumerados utilizando el intervalo de números naturales [0,11] en el cual 0 representa a enero, 1 febrero y así sucesivamente. De forma análoga se asignan los valores para días, comenzando por lunes hasta domingo, con valores comprendidos desde 0 al 6, respectivamente.

4.2.2. Detección y eliminación de *outliers*

La detección de *outliers* en el *set* de datos se realizó mediante el análisis de dendrograma. Éste se obtuvo aplicando *single-link clustering* en el *set* de datos, probando con tres métricas de distancia para su construcción: Manhattan, euclideana, y coseno. Debido a que se utiliza *single linkage* en la construcción del dendograma, se reconocerán como *outliers* aquellos registros que se incorporan al *cluster* global en último lugar, ya que corresponden a los mas alejados dentro del set. En la figura 4.10 se observan el dendrograma obtenido con las distancias Manhattan y euclideana[1], con las cuales se obtuvo el mismo resultado, mientras que en la figura 4.11 se observa el dendrograma con distancia coseno [2]. Además se eliminó el 1 % de los datos con mayor y menor área quemada.

[1] En la figura 4.10 (b) se se realiza un acercamiento al área en el que se encuentran los outliers
[2] En la figura 4.11(b) se se realiza un acercamiento al área en el que se encuentran los outliers

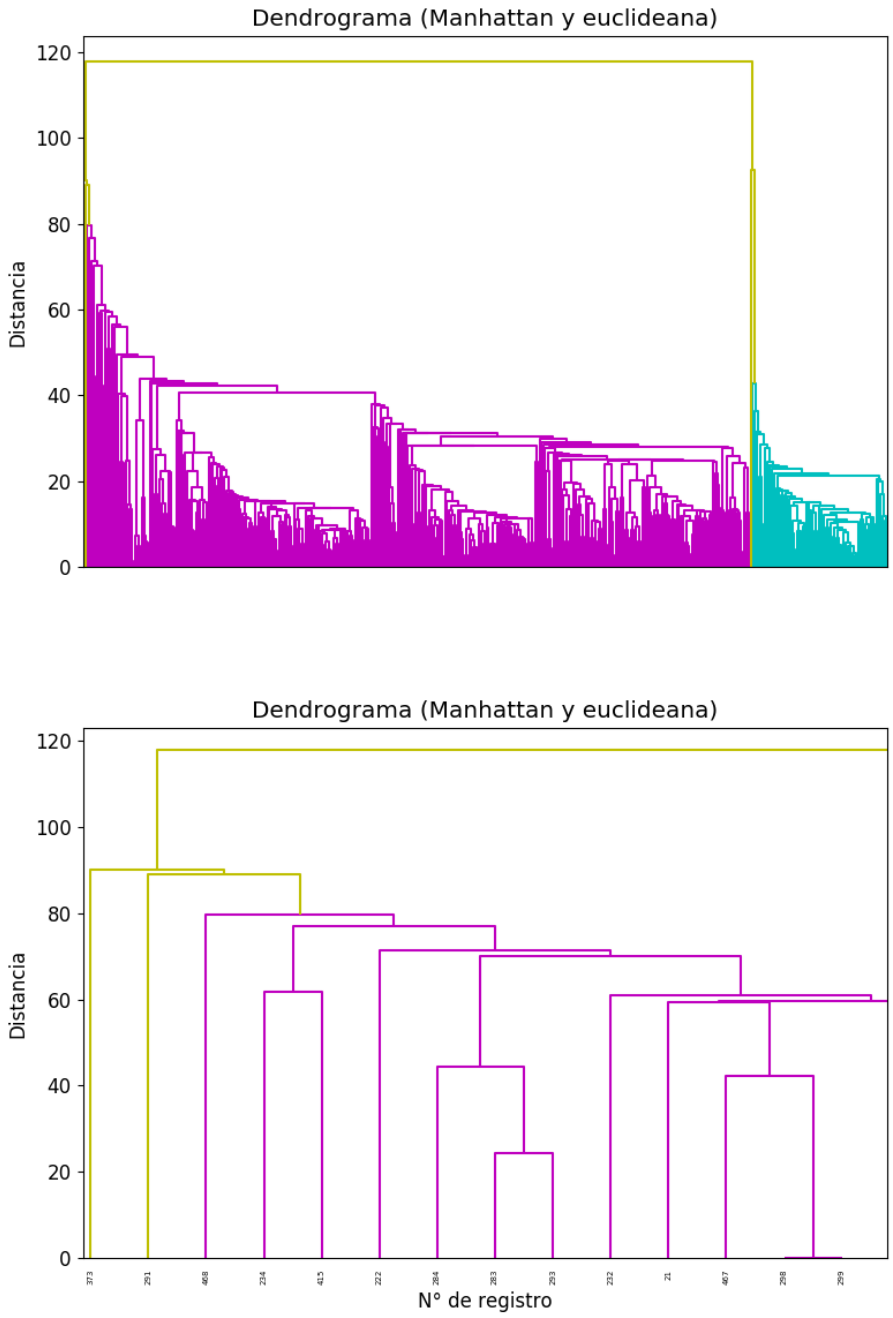


Figura 4.10: *Dendrograma obtenido con distancia Manhattan y euclideana (fabricación propia).*

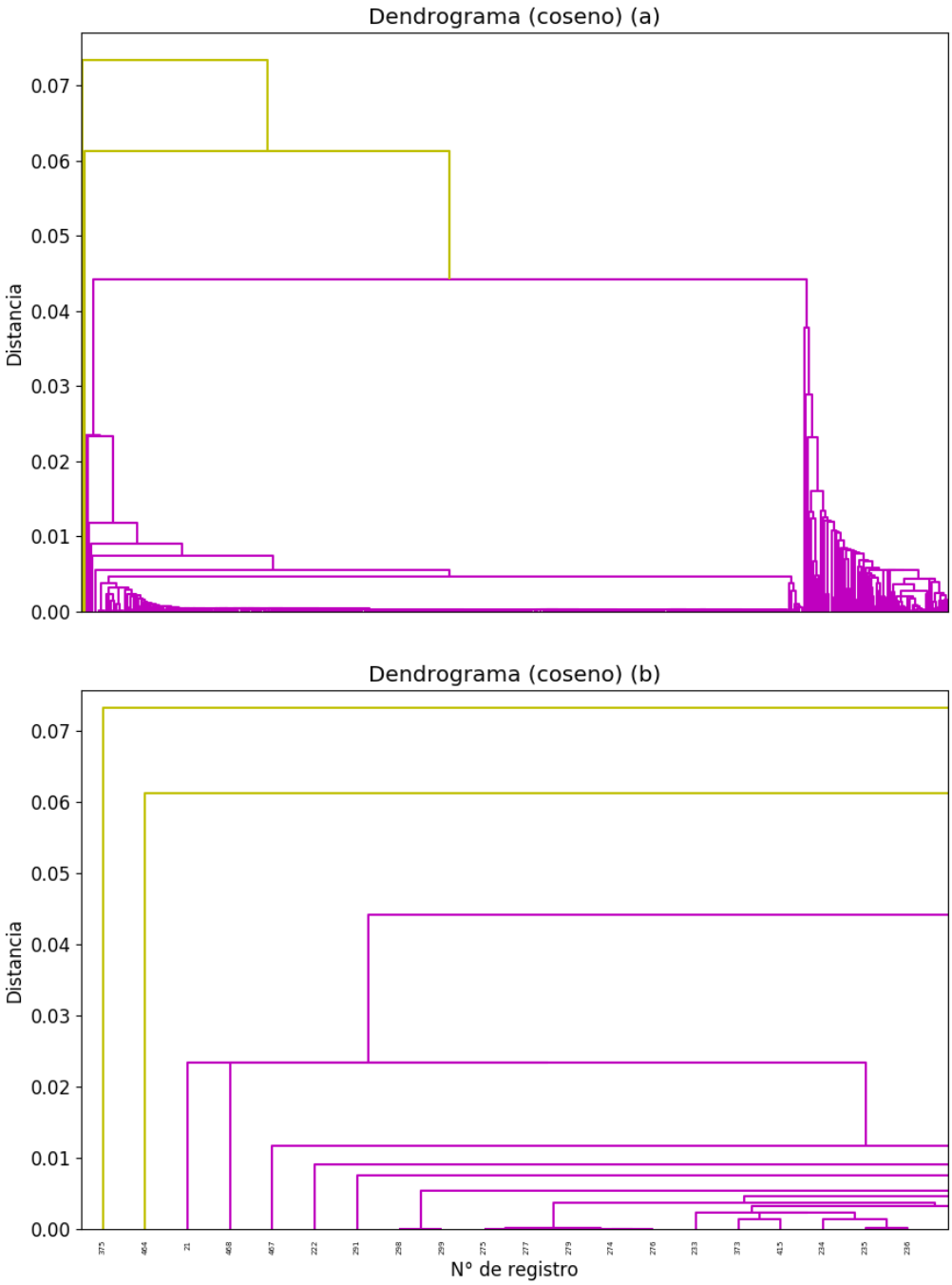


Figura 4.11: *Dendrogramas obtenido con distancia coseno (fabricación propia).*

4.2.3. Normalización

La normalización consiste en igualar las escalas de las características del *set* de datos dentro de un rango de valores determinado, técnica considerada fundamental en la aplicación de técnicas de *machine learning*, ya que evita que las variables de mayor escala tengan más impacto al calcular la salida del algoritmo. La normalización se realizará siguiendo las recomendaciones de [27], en el cual se recomienda el uso del *Min/Max Scaler* para establecer el rango de cada característica en el intervalo [0, 1].

4.2.4. Transformación de datos para cada tipo de Test

Como se dijo en el capítulo 3 para el desarrollo de la propuesta se realizan *tests* de regresión y clasificación, los cuales incluyen fases de preprocesamiento de datos que son excluyentes entre diferentes pruebas. A continuación se detallan las transformaciones realizadas exclusivamente para cada uno.

Test de regresión: Transformación logarítmica del área

Como se observó en la sección 4.1, la distribución del área tiene un sesgo en torno a 0. Para mejorar dicha distribución se aplica la transformación logarítmica del área mediante la ecuación 4.1, tal como se recomienda hacer en [7]. Los resultados de la transformación se observan en la figura 4.12. Los principales beneficios de esta transformación es que disminuye el rango de valores de los datos, resultando en una distribución en la que no existen *outliers* tan notorios como en el histograma de la figura 4.8.

$$f(area) = \log(area + 1) \tag{4.1}$$

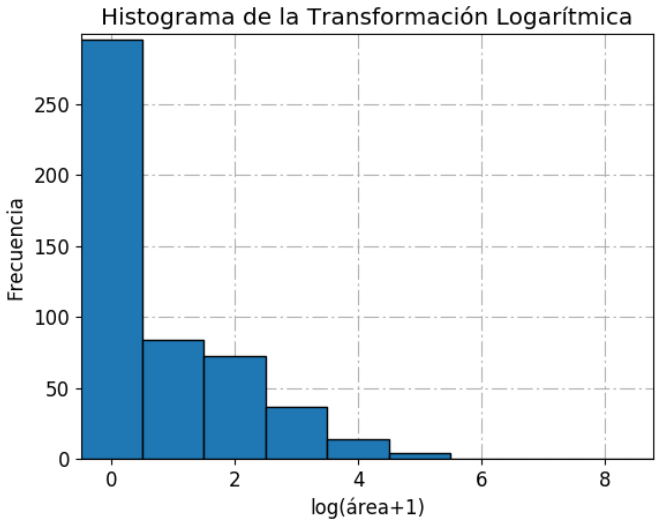


Figura 4.12: *Histograma de la Transformación Logarítmica del área (fabricación propia).*

Test de clasificación I

El problema de la predicción del área de incendios forestales fue propuesto originalmente en [7] como un problema de regresión. Sin embargo, en los estudios realizados posteriormente [8], [9] y [10] se demuestra que resolver el problema interpretándolo como uno de clasificación conlleva mejoras en la precisión alcanzada por los modelos. Esta transformación se logra reemplazando la característica *área*, por las etiquetas asignadas a los *clusters* generados al aplicar técnicas de *clustering* que indican cualitativamente el tamaño del incendio, la que se transforma en la nueva variable a predecir por el modelo. Cinco algoritmos de *clustering* fueron probados en el *set* de datos, siendo éstos: K-means, *Clustering* jerárquico con *linkage complete* y *Ward*, DBSCAN y *Spectral Clustering*.

Cada una de estas técnicas necesita la sintonización de parámetros exclusivos a cada algoritmo para su funcionamiento. Algoritmos como *K-means* y *Clustering* Jerárquico requieren como parámetro el valor de *k*, que indica el número de *clusters* que se desea encontrar; para determinar su valor se utilizó el análisis del *silhouette score*, métrica que indica la similitud de un objeto con el *cluster* al que fue asignado respecto del resto. Los valores de este *score*

se encuentran en el rango $[-1, 1]$, donde valores positivos indican que el registro se encuentra alejado de la frontera del *cluster*, valores cercanos a 0 indican que se encuentran en esta frontera y por último, valores negativos indican que el registro fue asignado erróneamente.

El análisis se realiza mediante la observación del gráfico generado por este *score*, donde el ancho de la silueta se asocia con el tamaño del cluster que representa. En la figura 4.13 se observan los distintos gráficos obtenidos con distintos valores de k. La figura 4.13 (a) no es un buen resultado, debido a que el *cluster* 0 supera significativamente en tamaño al *cluster* 1. Similar situación ocurre en la figura 4.13 (d) en la que existe una gran cantidad de registros asignados erróneamente. Por el contrario, al agrupar el *set* de datos en 3 o 4 *clusters* los registros se distribuyen equitativamente, situación reflejada en las figuras 4.13 (b) y 4.13 (c). Puesto que con 3 *clusters* se alcanza el mayor *silohuette score* promedio con un valor de 0.22, es que se decide establecer el valor de k en 3. Por otro lado, el algoritmo DBSCAN necesita la sintonización de los parámetros *eps* y *min points*, los cuales definen si un registro es considerado como un *core point* o como ruido. La elección del parámetro *eps* se realizó analizando el gráfico de los k-vecinos mas cercanos, donde se detecta el punto de inflexión para determinar el valor del *eps*. En el gráfico de la figura 4.14 se observa que el punto de inflexión ocurre en torno al punto 0.4, valor elegido como *eps*.

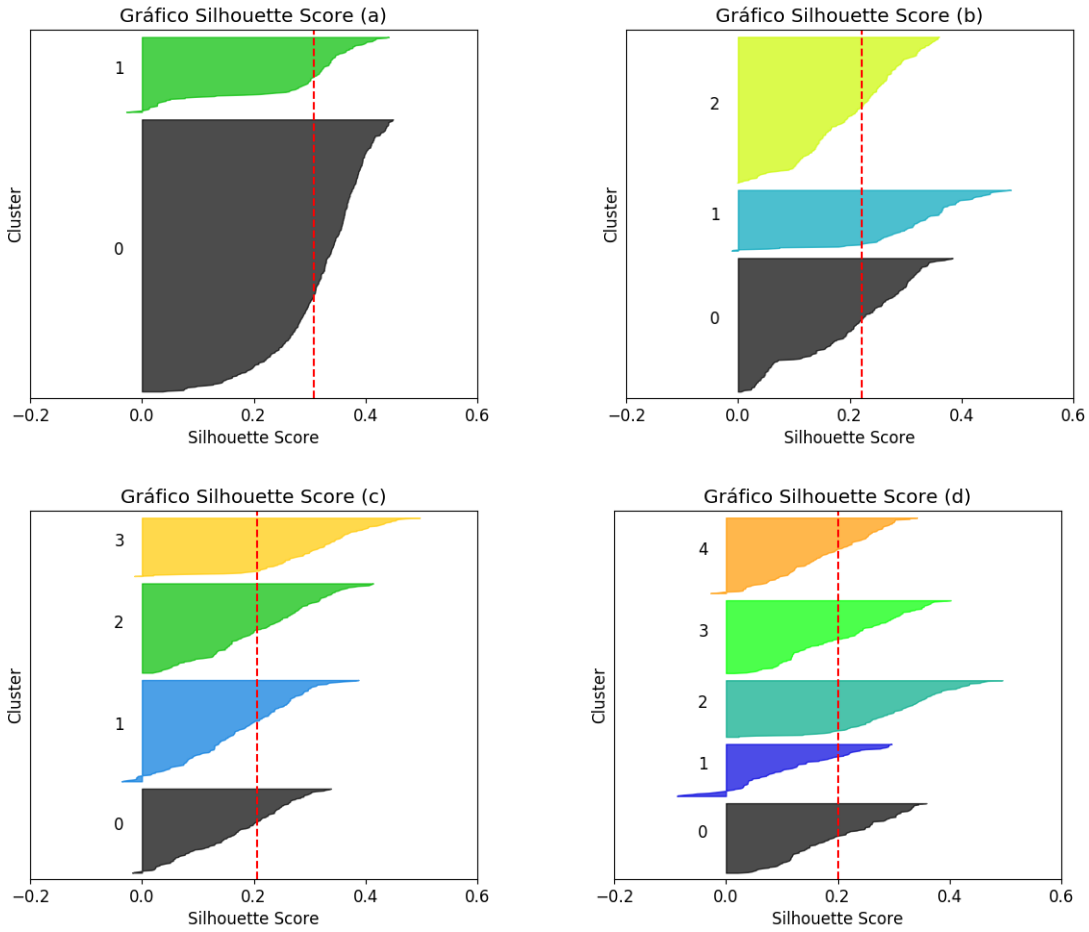


Figura 4.13: *Gráficos del Silohuette Score para valores de k iguales a 2 (a), 3(b), 4(c), 5(d) (fabricación propia).*

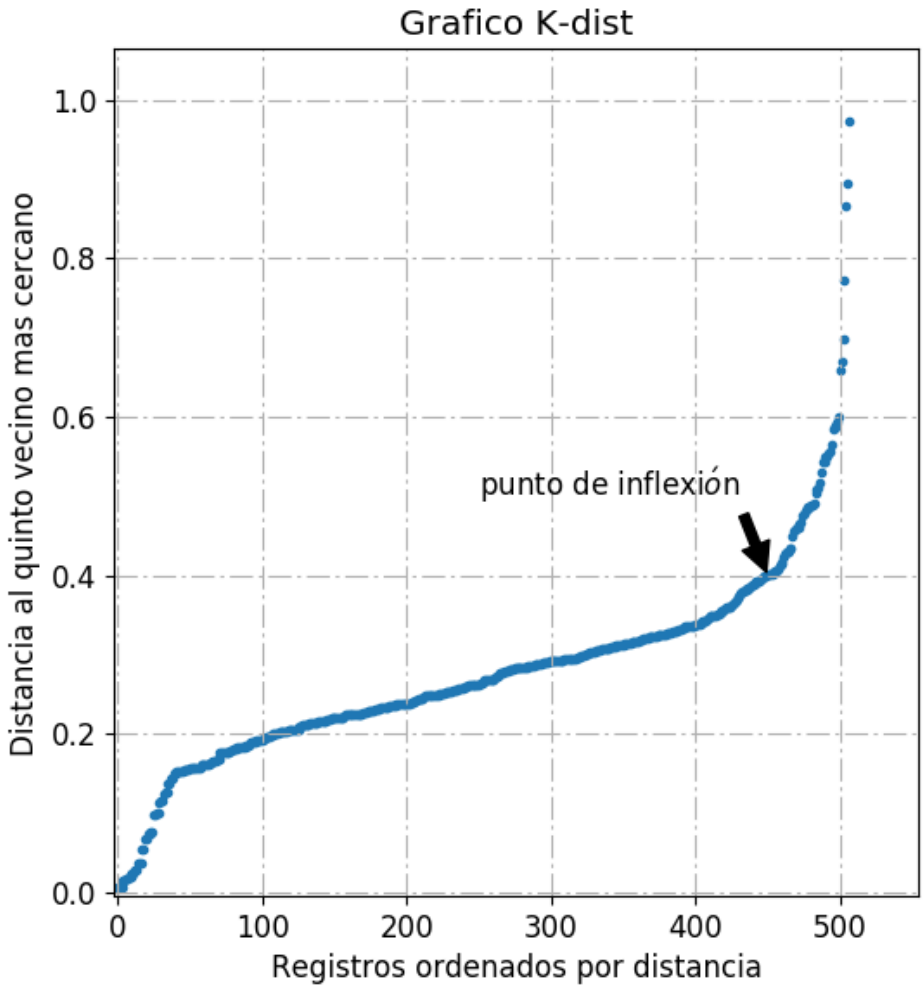


Figura 4.14: *Gráfico de la distancia del "quinto vecino mas cercano"para el set de datos. El punto de inflexión se encuentra cercano al valor 0.4, en el cual mas de 400 puntos se encuentran a una distancia igual o menor esta entre sí (fabricación propia).*

Test de clasificación II

Este *test* de clasificación sigue la misma lógica que el *test* de clasificación I, con la única diferencia que los valores del área son reemplazados con una etiqueta que representa el cuartil al que pertenece el área. Los cuartiles fueron calculados sin considerar los incendios

menores a $100m^2$, ya que al ser codificados con valor 0, provoca que los cuartiles 1 y 2 sólo contengan valores nulos. En el cuadro 4.4 se encuentran los rangos de los cuartiles juntos con las etiquetas generadas utilizando *Label Encoder*; a los registros cuya área es menor a 100 m^2 se les asigna la etiqueta 0.

Cuartil	Mínimo	Máximo	Etiqueta
1	0.09 [ha]	2.14 [ha]	1
2	2.14 [ha]	6.36 [ha]	2
3	6.36 [ha]	15.34 [ha]	3
4	15.34 [ha]	212.88 [ha]	4

Cuadro 4.4: *Cuartiles encontrados para el área (fabricación propia)*

4.2.5. Selección de atributos

La selección de atributos consiste en encontrar las características que mejor describen a la variable área de forma independiente, lo que se denomina como selección de atributos univariado [20]. Esto se llevó a cabo seleccionando los k mejores atributos obtenidos mediante la métrica de Información Mutua y los *tests* χ^2 y ANOVA.

4.3. Modelado y Evaluación

Finalizado el pre-procesamiento de datos se realiza el entrenamiento de cada SVM con el 70 % de los datos, y el 30 % restante se utilizó para evaluar la precisión de los modelos. La sintonización de hiper-parámetros y *kernels* de la SVM fue realizada mediante *Grid Seach* optimizado con *K-fold Cross Validation*, utilizando diez *folds* estratificados para el caso de los tests de clasificación. Los kernels y valores de C utilizados se encuentran en el cuadro 4.5, los valores de C fueron escogidos según las indicaciones entregadas en [27].

Parámetro	Valores
kernel	rbf, lineal, polinomial, sigmoide
C	1, 10, 100, 1000

Cuadro 4.5: *Conjunto de parámetros escogidos para entrenar cada tipo de test (fabricación propia)*

4.3.1. Evaluación

Test de Regresión

En el cuadro 4.6 se encuentra el detalle del mejor modelo encontrado vía *Grid Search*. El ranking de modelos fue ordenado en base al error medio absoluto (MAE) promedio obtenido al realizar su evaluación en los *test folds* durante la fase de Grid Search.

Parámetro	Valor
kernel	rbf
C	1000
Número de características	8
Características	X,Y, mes, día, ISI, temperatura, velocidad del viento, precipitaciones

Cuadro 4.6: *Configuración de la SVM para el test de regresión (fabricación propia).*

De este modo, la SVM con *kernel* de función de base radial es escogida como el mejor modelo de regresión, registrando un error medio absoluto de 10.24[ha] y una raíz del error cuadrático medio (RMSE) de 30.73. Por otra parte, la cantidad de registros correctamente predichos al aceptar 2[ha] de error fue de un 67.32 %, el cual fue calculado aplicando la función inversa de la transformación logarítmica en el valor precedido, tal como se realiza en [7]. Este modelo utiliza 8 características de las 12 disponibles en el *set* de datos, descartando

el uso de los 3 códigos del sistema FWI que indican la humedad de los combustibles presentes en el suelo. Por último, el tiempo de entrenamiento promedio fue de 7.68[ms] utilizando dicha configuración.

Test de clasificación I

Las etiquetas asignadas a cada *cluster* se encuentran en el cuadro 4.7 las cuales, se decidieron en base a los promedios del área encontrado en cada *cluster*, en donde al valor máximo se le asocia la etiqueta ”Grandeƶ al valor mínimo el valor ”Pequeño”. En el cuadro 4.8 se encuentran los resultados parciales obtenidos a través de este *test*, en el cual los modelos se organizan en base a la técnica utilizada en la fase de *clustering* con la que se realiza el reemplazo del área de una variable continua a categórica. Los resultados obtenidos mediante DBSCAN no son presentados, ya que bajo la sintonización encontrada a través del análisis del gráfico *K-dist* se encontró un único *cluster*.

Técnica de Clustering	***Cluster***	**Etiqueta**
HAC (complete)	0	Grande
	1	Mediano
	2	Pequeño
HAC (Ward)	0	Mediano
	1	Pequeño
	2	Grande

Técnica de Clustering	***Cluster***	**Etiqueta**
K-means	0	Grande
	1	Mediano
	2	Pequeño
Spectral Clustering	0	Grande
	1	Pequeño
	2	Mediano

Cuadro 4.7: *Etiquetas asignadas a cada cluster según técnica utilizada (fabricación propia).*

La configuración óptima de cada modelo se determinó mediante el ranking generado por *Grid Search*, el cual ordena en base a la precisión (*accuracy*) promedio registrada por los modelos en los *folds* de *testing* generados por el algoritmo de *k-fold cross validation*. A diferencia del *test* de regresión, en esta prueba se encontró más de un modelo con máxima precisión por cada técnica de *clustering*, en donde el mayor cantidad de modelos alternativos

Clustering	**Parámetros SVM**	**Número de características**	*Accuracy*	*Recall*
HAC (Complete)	(1) Kernel: sigmoide C : 100	8	0.99 (+/- 0.011)	0.99 (+/- 0.011)
	(2) Kernel: sigmoide C:100	9	0.95 (+/- 0.028)	0.95 (+/- 0.028)
HAC (Ward)	(3) Kernel: polinomial C: 100	7	0.95 (+/- 0.028)	0.95 (+/- 0.028)
K-means	(4) Kernel: polinomial C : 100	5	0.99 (+/-0.013)	0.99 (+/-0.013)
	(5) Kernel: sigmoide C: 1000	7		
Spectral Clustering	(6) Kernel: lineal C: 100 - 1000	4	0.98 (+/- 0.014)	0.98 (+/- 0.014)
	(7) Kernel: rbf - C: 100 (8) Kernel: polinomial - C: 1000 (9) Kernel: sigmoide - C: 1000	5		
	(10) Kernel: rbf - C: 100 (11) Kernel: polinomial - C: 1000	6		
	(12) Kernel: rbf C: 1000	9		
	(13) Kernel: rbf C: 1000	10		

Cuadro 4.8: *Resultados y configuración obtenido para cada técnica de clustering aplicado en el test de clasificación I (fabricación propia).*

se obtiene al utilizar *Spectral Clustering*. Las diferencias entre cada modelo yacen en los parámetros kernel y C utilizados en la SVM, tal como se observa en el cuadro 4.8, en el número de características utilizados para su entrenamiento resumidos en los cuadros 4.9 y 4.10 y por último en los tiempos promedio de entrenamiento registrados en el cuadro 4.11. La diferencia entre usar un *kernel* lineal y una función de base radial, sigmoide o polinomial se encuentra en la complejidad del modelo que será generado. Modelos complejos requieren, en este caso, de la sintonización de una mayor cantidad de hiperparámetros y además se arriesga a sobreajustar el modelo a los datos de entrenamiento, disminuyendo su capacidad de generalización.

Por otra parte, se observa una notoria diferencia entre la cantidad de características utilizadas para entrenar cada modelo. Sin embargo, hay un conjunto de características que son comunes a todos estos, las cuales son mes, día, temperatura y los códigos DMC y DC. Recordando que estas últimas, implícitamente incluyen las variables humedad relativa y precipitaciones.

Los modelos escogidos por cada técnica de *clustering* se encuentran en el cuadro 4.12. En el caso de *Spectral clustering* se escogió el modelo (6) dado que es un modelo menos complejo respecto de las otras alternativas, presentando un menor tiempo promedio de entrenamiento. Por otra parte, en las otras técnicas de *clustering* se encontró un conjunto más acotado de modelos de SVM por lo que la elección se realizó considerando este mismo tiempo, en conjunto con los modelos que requerían de la menor cantidad de características para ser entrenadas.

Test de Clasificación II

Los resultados obtenidos a través de este *test* se encuentran en el cuadro 4.13. Al igual que en el *test* anterior el ranking de modelos se ordenó en base a la precisión (*accuracy*) registrada por el modelo en los *folds* de *testing* generados en la fase de *Grid Search*. La SVM de *kernel* sigmoide es la que registra los peores resultados al evaluar su desempeño en el *set* de *testing* alcanzando un 46 % de precisión, además este modelo registra uno de los tiempos de entrenamiento más altos, necesitando en promedio 15[ms] en promedio.

Clustering	**Número de características**	**Características**
HAC (Complete)	(1) Kernel: sigmoide C : 100	Mes, Día, Temperatura, Humedad Relativa, Velocidad del viento, DMC, DC, ISI
	(2) Kernel: sigmoide C:100	Mes, Día, Temperatura, Humedad Relativa, Velocidad del viento, FFMC, DMC, DC, ISI
HAC (Ward)	(3) Kernel: polinomial C: 100	Mes, Día, Temperatura, FFMC, DMC, DC, ISI
K-means	(4) Kernel: polinomial C : 100	Mes, Día, Temperatura, DMC, DC
	(5) Kernel: sigmoide C: 1000	Mes, Día, Temperatura, Precipitaciones, DMC, DC, ISI

Cuadro 4.9: *Características utilizadas para la generación de los modelos de clasificación (fabricación propia).*

Clustering	**Número de características**	**Características**
Spectral Clustering	(6) Kernel: lineal C: 100 - 1000	Mes, Día, DMC, DC
	(7) Kernel: rbf - C: 100 (8) Kernel: polinomial - C: 1000 (9) Kernel: sigmoide - C: 1000	Mes, Día, Temperatura, DMC, DC
	(10) Kernel: rbf - C: 100 (11) Kernel: polinomial - C: 1000	Mes, Día, Temperatura, DMC, DC, ISI
	(12) Kernel: rbf C: 1000	Mes, Día, Temperatura, Humedad Relativa, Velocidad del viento, DMC, DC, ISI
	(13) Kernel: rbf C: 1000	Mes, Día, Temperatura, Humedad Relativa, Velocidad del viento, Precipitaciones, FFMC, DMC, DC, ISI

Cuadro 4.10: *Características utilizadas para la generación de los modelos de clasificación (fabricación propia).*

Clustering	**Parámetros SVM**	**Número de características**	**Tiempo entrenamiento promedio [ms]**
HAC (Complete)	(1) Kernel: sigmoide C : 100	8	4.00 (+/- 1.10)
	(2) Kernel: sigmoide C:100	9	3.70 (+/- 1.00)
HAC (Ward)	(3) Kernel: polinomial C: 100	7	4.30 (+/- 1.27)
K-means	(4) Kernel: polinomial C : 100	5	3.30 (+/- 1.19)
	(5) Kernel: sigmoide C: 1000	7	4.30 (+/- 1.27)
Spectral Clustering	(6) Kernel: lineal C: 100 - 1000	4	3.10 (+/- 0.94) 4.00 (+/- 1.41)
	(7) Kernel: rbf - C: 100 (8) Kernel: polinomial - C: 1000 (9) Kernel: sigmoide - C: 1000	5	4.00 (+/- 1.27) 3.90 (+/- 0.94) 3.60 (+/- 0.92)
	(10) Kernel: rbf - C: 100 (11) Kernel: polinomial - C: 1000	6	4.00 (+/- 0.89) 4.30 (+/- 1.35)
	(12) Kernel: rbf C: 1000	9	5.10 (+/-1.14)
	(13) Kernel: rbf C: 1000	10	5.50 (+/-2.06)

Cuadro 4.11: *Tiempos de entrenamiento registrados para cada modelo originado en el test de clasificación I (fabricación propia).*

Clustering	**Parametros SVM**	**N parámetros**	***Test Accuracy***
HAC (Complete)	Kernel: sigmoide C : 100	8	0.97
HAC (Ward)	Kernel: polinomial C: 100	7	0.95
K-means	Kernel: polinomial C : 100	5	0.98
Spectral Clustering	Kernel: lineal C: 100 - 1000	4	0.98

Cuadro 4.12: *Precisión alcanzada en el set de testing por los mejores modelos encontrados en el test de clasificación I (fabricación propia).*

Característica	**Valor**
Parámetros SVM	C: 100 Kernel: sigmoide
N de características	10
Características	X, Y, Mes, Día, FFMC, DC, ISI, temperatura, velocidad del viento, precipitaciones
Test Accuracy	0.46
Tiempo de entrenamiento[ms]	15

Cuadro 4.13: *Resultados y configuración de la SVM para el test de clasificación II (fabricación propia).*

4.3.2. Conclusiones del Experimento

En el desarrollo de esta propuesta se evaluó el desempeño de tres modelos de predicción del tamaño de incendios forestales a partir de dos métodos distintos. Uno, en el cual se resuelve el problema como originalmente fue propuesto, es decir, mediante la predicción cuantitativa del área y por otra parte, reformulando dicho problema a uno de clasificación, en el cual se predice cualitativamente el tamaño de los incendios forestales a través de etiquetas. En el primero de los *tests*, se resolvió el problema empleando una SVM para realizar la regresión del área quemada en hectáreas, obteniendo mejores resultados que los alcanzados en el estudio [7], en el cual se logró predecir correctamente el área del 61 % de los registros, aceptando 2[ha] de error de predicción. Mientras que el modelo propuesto en esta memoria, aumenta dicho valor en aproximadamente 6 puntos porcentuales, llegando a un 67.32 %. En la figura 4.15 además se observa la disminución de las métricas MAE y RMSE en comparación con este mismo estudio.

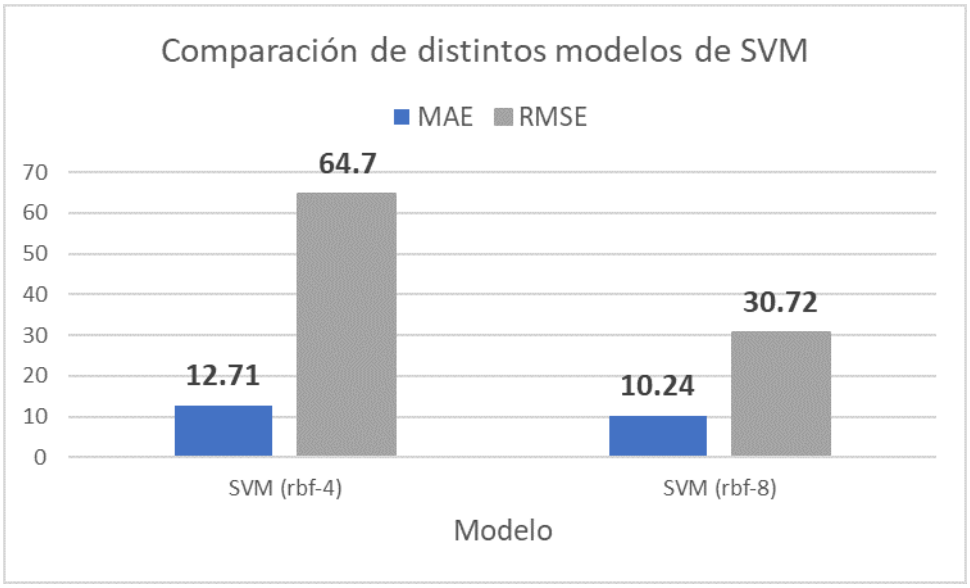


Figura 4.15: *Comparación de los errores registrados por el modelo propuesto en [7] (SVM rbf-4) y el propuesto en este memoria (SVM rbf-8) (fabricación propia).*

Sin embargo, esta mejora del nivel de predicción tiene como consecuencia un aumento de la cantidad de datos necesarios, ya que el modelo propuesto en esta memoria requiere de 8

características del *set* de datos original, mientras que el modelo propuesto en [7] sólo utiliza las 4 variables meteorológicas.

Por otra parte, dos modelos se propusieron para resolver el problema de clasificación, los cuales registraron resultados totalmente opuestos. Para la construcción del primero de ellos, se probaron diferentes técnicas de *clustering* en la fase de transformación de datos, con el objetivo de generar etiquetas que describieran el tamaño del incendio forestal. Cinco técnicas fueron probadas, de las cuales cuatro tuvieron éxito al detectar los *clusters* en el *set* de datos. El principal inconveniente que presentan dichas técnicas, es la búsqueda de los parámetros correctos para cada algoritmo, que en el caso de *K-means*, *Clustering* jerárquico y *Spectral Clustering* es el valor de k, mientras que en DBSCAN son los valores de *min points* y *eps*. A pesar de utilizar los métodos recomendados para la elección de dichos parámetros, no se obtuvieron los resultados esperados mediante DBSCAN; sin embargo, al realizar el mismo análisis en una proyección en dos dimensiones de los datos utilizados el algoritmo logra detectar correctamente *clusters*, lo cual indicaría que dicha técnica está mejor situada para casos de baja dimensionalidad. Los mejores modelos se obtienen al aplicar las técnicas de *K-means* y *Clustering* Jerárquico con *complete linkage* en la fase de transformación de datos, alcanzando una precisión de un 99 %, superando el 91 % obtenido en [10] y la precisión alcanzada por las redes neuronales del estudio [9]. Este aumento de la precisión, al igual que en el *test* de regresión, requiere de una mayor cantidad de datos en comparación con los modelos propuestos en dichos estudios, que al igual que en [7] solo utilizan los atributos meteorológicos del *set* de datos para realizar la predicción. Sin embargo, la elección de este subconjunto en dichos estudios es realizada de forma arbitraria, sin utilizar una técnica de selección de parámetros como las implementadas en el desarrollo de esta memoria.

El *test* de clasificación II alcanzó solo un 46 % precisión lo cual se considera deficiente en comparación con los resultados obtenidos por los otros modelos propuestos en esta memoria, y con los registrados en los estudios [7], [8], [9] y [10]. Su aplicación no se recomienda en *sets* de datos cuyas áreas presenten un sesgo tan notorio como el utilizado en este memoria, puesto que origina una desproporción en los rangos encontrados en el último cuartil, en donde se abarcan valores desde los 15[ha] hasta las 212[ha].

En cuanto al proceso de transformación de datos, se observó que la eliminación del 1 % de

los datos y la detección de *outliers* via dendrograma tuvo un impacto positivo en el *test* de regresión y en la implementación de las técnicas de *clustering*. Sobre todo en estas últimas, debido a que técnicas como *K-means* no son robustas ante la presencia de este tipo de datos. Por otra parte, se observó que la selección de los mejores atributos para la generación del modelo no se vio afectada por las métricas utilizadas para realizar el ranking, puesto que se dieron casos en que modelos encontrados por *Grid Search* con la misma precisión y configuraciones sólo se diferenciaban en la métrica utilizada para la elección de las características, las cuales entregaban el mismo subconjunto de ellas.

Finalmente se concluye que ambos modelos, tanto el de regresión como el de clasificación I, presentan resultados que permiten evaluar positivamente su labor predictiva. Sin embargo, ambos modelos presentan ventajas y desventajas que se deben considerar al momento de decidir implementar una técnica por sobre la otra. La principal ventaja de usar un modelo de regresión por sobre uno de clasificación yace en que el primero permite una interpretación directa de la salida del modelo ya que se obtiene un valor numérico aproximado del área quemada. Sin embargo la presencia de *outliers* en el *set* de datos empeora la calidad del modelo, por lo que se sugiere la implementación de una técnica que permita su detección en la fase de preparación de datos. Por otra parte, el modelo de clasificación I presenta la principal ventaja de contar con mejores niveles de precisión, a cambio de implementar una técnica de *clustering* en la fase de transformación de datos. En este caso no se recomienda utilizar una técnica de *clustering* basada en densidad, dados los deficientes resultados que DBSCAN presentó al aplicarse en el set de datos. Otro factor importante que se debe tener en consideración al momento de implementar este tipo de técnicas es que se debe contar con el conocimiento experto necesario para poder realizar un etiquetado de *clusters* que permita una interpretación objetiva de la salida del modelo.

Conclusiones

En el desarrollo de esta memoria se abordó la problemática de la predicción del área quemada por incendios forestales, temática que ha cobrado gran importancia en Chile dado los últimos acontecimientos registrados en las meses de verano a lo largo del país. Utilizando la metodología de CRISP-DM se elaboraron tres modelos predictivos planteado bajo dos perspectivas: como un problema de regresión y por otra parte, como uno de clasificación. En general el desarrollo de la propuesta fue fluido, debido a que el modelo utilizado cuenta con un alto nivel de descripción, lo cual facilita estructurar el proyecto de forma expedita. El desarrollo de esta memoria resultó ser una enriquecedora experiencia para comprender el desarrollo de un proyecto de minería de datos y tener una experiencia cercana a lo que sería su ejecución en un ambiente de negocios.

Tres modelos fueron propuestos en esta memoria, de los cuales dos de ellos presentaron una mejora con respecto a los resultados obtenidos en estudios similares, siendo estos los modelos de regresión y de clasificación por *clustering*. Ambos, representan diferentes perspectivas de resolución del problema, cada uno con sus respectivas ventajas y desventajas. Las principales ventajas de ambos modelos yacen en su alto nivel de precisión, el cual beneficia en mayor manera al modelo de regresión, al aumentar su precisión en un 6 % con respecto a resultados obtenidos en estudios como [7], a costa de requerir una mayor cantidad de características para la formulación del modelo, y de la aplicación de métodos de transformación de datos para que éstos puedan ser procesados por la SVM. Sin embargo, se debe considerar que bibliotecas como *scikit learn* contiene implementaciones para realizar dichas transformaciones de manera sencilla y eficiente. Por otra parte, el modelo de regresión cuenta con

una salida que permite interpretar directamente los resultados; en cambio, el modelo de clasificación I, requiere de un análisis de los *clusters* encontrados para generar etiquetas que describan correctamente el tamaño de los incendios agrupados en dicho conjunto de datos. Además, este último test requiere de la sintonización adicional de los hiperparámetros de la técnica de *clustering* que se desea utilizar. Otra diferencia que se da entre ambos modelos corresponde a la cantidad de características utilizadas para su entrenamiento, el modelo de regresión ocupa sólo ocho características del *set* de datos, mientras que en en una de sus configuraciones el del modelo de clasificación sólo se utilizan cuatro, lo cual tiene un impacto menor en el tiempo de entrenamiento, alcanzando una diferencia de 1[ms] aproximadamente. Finalmente la aplicación del modelo de clasificación II queda descartada, dado el deficiente desempeño alcanzado, tanto en niveles de precisión como en tiempos de entrenamiento.

En el desarrollo de la memoria se logró cumplir con el objetivo principal, el cual correspondía a contribuir con un nuevo estudio para la predicción de incendios forestales, que además puede ser utilizado como un punto de referencia para la ejecución de un proyecto de minería de datos que resuelvan otro tipo de temáticas. La labor que resultó ser más compleja fue determinar los parámetros óptimos de la SVM, en particular en el *test* de clasificación I, ya que se obtuvieron una cantidad considerable de modelos de igual precisión, con pequeñas diferencias en la configuración de la SVM, los cuales aumentaban de manera considerable al agregar más parámetros de búsqueda para *Grid Search*, lo cual además, impactaba directamente en el tiempo de ejecución. Es por esto, que se valora la existencia de estudios como los de [27], en los que se entrega una guía para aquellos que cuentan con menos experiencia en el desarrollo de modelos predictivos basados en el uso de SVMs y libros como [20] en los que se ejemplifica el uso de la bibliteca *scikit-learn* de python y como utilizar correctamente sus funciones. Otro aspecto que fue complejo en el desarrollo de la memoria fue la obtención del *set* de datos, puesto que no existe una gran cantidad de ellos que se encuentren públicos, sobre todo considerando que los sitios web de la entidades encargada del combate y control de incendios forestales cuentan con estadísticas históricas de la ocurrencia, así como también informes en tiempo real de la ocurrencia de incendios forestales y las condiciones del meteorológicas.

Como trabajo futuro se proponen la aplicación de estos modelos con un *set* de datos obtenidos en territorio nacional, ya que son distintas las condiciones climáticas que afectan el país; de este modo se podría concluir si existen similitudes entre las variables que afectan la ignición y propagación de incendios forestales con otros países. Lamentablemente no se logró encontrar un *set* de datos público del cual se pudiera hacer uso ya que, a la fecha de presentación de esta memoria, sólo se cuentan con estadísticas a nivel nacional de las ocurrencias de estos siniestros en el país. Por otra parte, sería interesante implementar un método de visualización de los resultados obtenidos a través del modelo, ya sea a través de una aplicación web, en la cual se pueda indicar la zona en la que se ubica el incendio, el tamaño probable de éste, y recomendaciones sobre las acciones a tomar por los expertos.

Finalmente, el desarrollo de la memoria fue una grata experiencia, puesto que constituyó un primer acercamiento al desarrollo profesional de un modelo predictivo, guiado a través de un proceso utilizado ampliamente en la industria CRISP-DM. Sin duda, la formación profesional entregada por la universidad constituye un base sólida de conocimiento, destacando asignaturas electivas como Máquinas de Aprendizaje, en el cual se estudiaron las bases de lo que es el aprendizaje supervisado, Reconocimiento de Patrones en Mineria de Datos en el cual se abarcan, entre otra cosas, la teoría de las técnicas utilizadas para realizar *clustering* de datos; y finalmente Bases Técnologicas para la Inteligencia de Negocios el cual me permitió conocer y descubrir mi interés por esta área de la informática.

Bibliografía

[1] CONAF, Incendios Forestales en Chile [en línea] <http://www.conaf.cl/incendios-forestales/incendios-forestales-en-chile/> [consulta: 27 Julio 2017]

[2] CONAF Quiénes Somos [en línea] <http://www.conaf.cl/quienes-somos/> [consulta: 27 Julio 2017]

[3] Mónica Garrido V. Incendios en Chile: El cuarto más devastador de los últimos 15 años en el mundo [en línea] La Tercera. 4 de febrero, 2017 <http://www.latercera.com/noticia/incendios-chile-los-cuartos-mas-devastadores-los-ultimos-15-anos-mundo/> [consulta: 27 Julio 2017]

[4] Natural Resources Canada, Forest Fires [en línea] <http://www.nrcan.gc.ca/forests/fire-insects-disturbances/fire/13143> [consulta: 27 Julio 2017]

[5] United States Department of Agriculture. The rising cost of wildfire operations [en línea] https://www.fs.fed.us/sites/default/files/2015-Fire-Budget-Report.pdf [consulta: 27 Julio 2017]

[6] Veronica Marín. Más de $350 mil millones ha gastado el Gobierno de Bachelet en el combate de desastres naturales [en línea] Emol. 7 de Junio, 2017 <http://www.emol.com/noticias/Nacional/2017/06/07/861647/Mas-de-350-mil-millones-ha-gastado-el-Gobierno-de-Bachelet-en-el-combate-de-desastres-naturales.html> [consulta: 27 Julio 2017]

[7] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data," pp. 1 – 12, 2007.

[8] R. D. H. M. K. S. Yong Poh Yu, Rosli Omar and A. R. Nik, "Pattern clustering of forest fires based on meteorological variables and its classification using hybrid data mining methods," Journal of Computational Biology and Bioinformatics Research, pp. 47 – 52, 2011.

[9] A. M. "Estimation of the burned area in forest fires using computational intelligence techniques," Procedia Computer Science 12, pp. 282 – 287, 2012.

[10] K. M. Guruh Fajar Shidik, "Predicting size of forest fire using hybrid model," pp. 316 – 327, 2014.

[11] SAS Data Mining From A to Z: How to Discover Insights and Drive Better Opportunities [en línea] <https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-mining-from-a-z-104937.pdf> [consulta: 20 agosto 2017]

[12] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth Crisp-DM 1.0 Step-by step data mining guide [en línea] <https://www.the-modeling-agency.com/crisp-dm.pdf> [consulta: 20 agosto 2017]

[13] SAS® Documentation Getting Started with SAS® Entreprise Miner™ 14.2[en línea] <http://support.sas.com/documentation/cdl/en/emgsj/70152/PDF/default/emgsj.pdf> [consulta: 20 agosto 2017]

[14] Knowledge Discovery in Databases [en línea] <https://www.cise.ufl.edu/ ddd/cap6635/Fall-97/Short-papers/KDD3.htm> [consulta: 20 agosto 2017]

[15] Overview of the KDD Process [en línea] <http://www2.cs.uregina.ca/ dbd/cs831/notes/kdd/1_kdd.html > [consulta : 19 octubre 2017]

[16] CONGRESO Argentino de Ciencias de la Computación (14ª, 2011, La Plata, Argentina). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. La Plata, 2011.

[17] Umair Shafique, Haseeb Qaiser. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research. 12(1):217-222, Noviembre, 2014

[18] CONFERENCIA IADIS European Conference on Data Mining (2008, Amsterdam, Países Bajos). KDD, semma and CRISP-DM: A parallel overview. Amsterdam, 2008.

[19] Jiawei Han, Micheline Kamber, Jian Pe. Data Mining Concepts and Techniques, 3 ª ed. Morgan Kaupmann. 2011. 744p

[20] Robert Layton. Learning Data Mining with Python. 1 ª ed. Birmingham, Packt Publishing, 2015. 347p.

[21] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Cluster Analysis: Basic Concepts and Algorithms [en línea] <http://www-users.cs.umn.edu/ kumar/dmbook/index.php > [consulta: 22 agosto 2017]

[22] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. 482p.

[23] Trevor Hastie Robert Tibshirani Jerome Friedman. The Elements of Statistical Learning Data Mining,Inference,and Prediction. 2ª ed. Estados Unidos, Springer, 2001. 745 p.

[24] Enrique J. Carmona Suárez, Tutorial sobre Máquinas de Vectores Soporte (SVM), pp. 1 - 25. 2013

[25] World Climate & Climate Information [en línea] <https://weather-and-climate.com/average-monthly-min-max-Temperature-fahrenheit,braga,Portugal > [consulta: 19 octubre 2017]

[26] Center for Machine Learning and Intelligent Systems [en línea] <https://archive.ics.uci.edu/ml/datasets/Forest+Fires > [consulta: 19 octubre 2017]

[27] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A Practical Guide to Support Vector Classification pp. 1 – 16, 2008

[28] Alberta Wildfire [en línea] <http://wildfire.alberta.ca/wildfire-status/fire-weather/understanding-fire-weather.aspx > [consulta: 19 octubre 2017]

[29] Agencia Estatal Boletín Oficial del Estado [en línea] <https://www.boe.es/buscar/act.php?id=BOE-A-2014-11493 > [consulta: 19 octubre 2017]