PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

MÁQUINAS VECTORIALES HÍBRIDAS PARA CLASIFICAR ACCIDENTES DE TRÁNSITO EN LA REGIÓN METROPOLITANA

MARCELO NICOLÁS FARÍAS CONCHA

TESIS DE GRADO
MAGISTER EN INGENIERÍA INFORMÁTICA

DICIEMBRE 2011

Pontificia Universidad Católica de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

MÁQUINAS VECTORIALES HÍBRIDAS PARA CLASIFICAR ACCIDENTES DE TRÁNSITO EN LA REGIÓN METROPOLITANA

MARCELO NICOLÁS FARÍAS CONCHA

Director de Tesis: **Nibaldo Rodríguez Agurto**

Programa: **Magister en Ingeniería Informática**

DICIEMBRE 2011

Dedicatoria

A mi familia y amigos, en especial a mis padres que estuvieron siempre para darme todo su apoyo y comprensión en este gran proceso para convertirme en un profesional.

Índice de Contenidos.

Lista de Figuras.

Lista de Tablas.

Abreviaciones y Siglas.

- OMS: Organización mundial de la salud.
- CONASET: Comisión nacional de seguridad de tránsito.
- SVM: Support vector machine.
- SV: Support vector.
- LS-SVM: Least square support vector machine.
- PSO: Particle swarm optimization.
- QPSO: Quantum particle swarm optimization.
- IPSO: Improved particle swarm optimization.
- ERM: Empirical Risk Minimization.
- SRM: Structural risk minimization.
- QP: Quadratic programation.

Resumen

Debido al aumento en los accidentes de tránsito, se propone realizar un estudio para clasificar el estado de las personas involucradas, específicamente para accidentes registrados en la región Metropolitana. Entonces, para realizar la clasificación de los accidentes de tránsito fueron utilizadas Máquinas de Soporte Vectorial (SVM), herramienta que dado un conjunto de muestras como ejemplo para entrenamiento permite etiquetar las clases y con esto entrenar la SVM para construir un modelo que prediga la clase de una nueva muestra. Esta técnica a pesar de ser robusta, también posee debilidades, las que se presentan como un problema combinatorial en la estimación y ajuste de sus parámetros de entrada. La obtención de buenos resultados depende de las características intrínsecas que presentan las SVM, además de la correcta elección de la función kernel y de los parámetros de entrada. La elección y ajuste de los parámetros fue realizada con un algoritmo evolutivo de Optimización por Enjambres de Partículas (PSO). Finalmente, para resolver el problema se desarrollaron distintos modelos utilizando SVM con algoritmos PSO, con lo que se buscaba clasificar el grado de severidad con el que resultan las personas involucradas en los accidentes de tránsito, ilesos o lesionados. En la búsqueda de mejores resultados también se utilizaron variaciones de PSO, generando distintos modelos, comparando los resultados obtenidos y con esto poder realizar la mejor elección para obtener óptimos resultados en la clasificación. Dado lo anterior, el mejor resultado se obtuvo para la comuna de Puente Alto, con un 94% de exactitud, 100% de sensibilidad y 83% de especificidad.

Palabras claves: Clasificación, Máquinas de Soporte Vectorial (SVM), Optimización por Enjambre de Partículas (PSO), Accidentes de Tránsito.

Abstract

Due to increased traffic accidents, a study is proposed to classify the status of those involved, especially for accidents in the Metropolitan region. Then, for the classification of traffic accidents were used Support Vector Machines (SVM), a tool that given a set of training samples as examples allows you to tag classes and thus train the SVM to build a model that predicts the class of a new sample. This technique despite being robust, it also has weaknesses, which are presented as a combinatorial problem in estimating and adjusting their input parameters. Obtaining good results depends on the intrinsic characteristics presented by SVM also the correct choice of the kernel function and the input parameters. The choice and adjustment of parameters was performed with an evolutionary algorithm of Particle Swarm Optimization (PSO). Finally, to solve the problem different models were developed used SVM with PSO algorithms, which sought to classify the degree of severity of the people who are involved in traffic accidents, this can be uninjured or injured. Searching better results, variations of PSO where used, generating different models, comparing the results obtained with this to make the best choice for optimal results in the classification. Therefore, the best results were obtained for Puente Alto, with 94% accuracy, 100% sensitivity and 83% specificity.

Keywords: Classification, Support Vector Machine (SVM), Particle Swarm Optimization (PSO), Traffic Accidents.

1. Introducción.

Los siniestros de tránsito son eventos complejos y aleatorios que involucran una variedad de factores que se conjugan para su ocurrencia, entre las que destacan los factores humanos, del entorno, del estado del vehículo y del tránsito, entre otros. Los expertos [1] también coinciden en que estos siniestros ocurren en gran medida porque no se respetan las reglamentaciones y normas existentes.

Los accidentes de tránsito a nivel mundial constituyen uno de los principales problemas sociales que han surgido en los últimos años. Además, esta problemática ha ido aumentando considerablemente su protagonismo y se ha convertido en una de las principales causas de mortalidad dentro de los distintos grupos etarios según la Organización Mundial de la Salud (OMS) [1]. La proyección que se maneja de su influencia dentro de la salud de la población es bastante importante, puesto que pasa de estar dentro de las diez primeras causas de mortalidad a nivel mundial en el año 2004, a ubicarse entre las cinco de mayor relevancia para el año 2030.

Estos accidentes además de la implicancia que tienen sobre las familias de los involucrados, influyen directamente en un alza considerable del gasto público y privado asociado al gran despliegue que se produce.

Esta situación como es de conocimiento de todos se repite claramente en Chile tal como lo confirma la Comisión Nacional de Seguridad de Tránsito (CONASET) [2]. Además, el motivo por el cual el estudio se realiza en la región Metropolitana, es porque en ésta se encuentra concentrada la mayor población del país, lo que trae como consecuencia que también se encuentre el mayor parque automotriz.

Clasificar los distintos tipos y datos presentes en los accidentes de tránsito resulta de gran ayuda al momento de determinar las principales causas de estos eventos. Esta información podría ser utilizada por las autoridades y determinar un conjunto de acciones a tomar para prevenir y mitigar la cantidad de accidentes y/o la gravedad con que resultan las personas involucradas en éstos. Otra utilidad que pretende prestar este trabajo es la reducción de las grandes sumas de dinero que deben desembolsarse a causa de estos accidentes.

Como se menciona anteriormente, se busca clasificar los accidentes de tránsito ocurridos en la región Metropolitana mediante la utilización de Máquinas de Soporte Vectorial con algoritmos evolutivos, puesto que estas herramientas presentan buenas características y también resultan ser consistentes para realizar este tipo de tareas tomando en cuenta que ya existen técnicas mejoradas que servirán para este propósito. Entonces se busca un modelo que defina el problema y con esto dar una solución satisfactoria.

1.1 Objetivo General.

Desarrollar un modelo de clasificación de datos para accidentes de tránsito de la región Metropolitana, utilizando Máquinas de Soporte Vectorial (SVM) con algoritmos de Optimización por Enjambre de Partículas (PSO) y algunas variaciones de éste.

1.2 Objetivos Específicos.

- Explicar la funcionalidad de las Máquinas de Soporte Vectorial (SVM) y de la optimización por enjambre de partículas (PSO).
- Diseñar la estructura y estimar los parámetros del Clasificador Vectorial de Soporte (CVS) utilizando PSO.
- Evaluar el porcentaje de exactitud del Clasificador Vectorial de Soporte (CVS).

1.3 Organización del Texto.

El trabajo se encuentra organizado de la siguiente manera, en el capítulo 2 se desarrolla la teoría y la técnica de las Máquinas de Soporte Vectorial (SVM), en la sección 2.1 se describen las SVM específicamente para problemas de Clasificación, luego en 2.2 se explican las funciones Kernel, y se detallan cuales son los más utilizados, en 2.3 se presentan los Multiplicadores de Lagrange y la solución para el problema de optimización que aquí se presenta, en 2.4 se expone la variante de Máquinas de Soporte Vectorial de Mínimos Cuadrados que se utilizará como base del desarrollo del trabajo.

En el capítulo 3 se introduce a la técnica de Optimización por Enjambre de Partículas y se especifica su aporte en el estudio, luego en 3.1 se presentan algunas de sus variantes, las que permiten realizar la búsqueda de los parámetros óptimos del modelo que se construye.

En el capítulo 4 se detalla el Clasificador de Soporte Vectorial, en 4.1 se muestra el modelo general que será origina para realizar el estudio, en 4.2 se muestran los parámetros de entrada que se utilizarán y en 4.3 están las métricas con las cuales será evaluado el Clasificador de Soporte Vectorial.

En el capítulo 5 se explican y representan los datos que se utilizarán, y en la sección 5.1 se encuentra la estructura original de los datos.

En el capítulo 6 se expone el desarrollo Clasificador de Soporte Vectorial, en 6.1 el proceso de preparación de los datos, en 6.2 se especifican los datos que serán utilizados, en 6.3 se presenta la función de costo que será utilizada, mientras que en 6.4 se detalla el modelo de Máquinas de Soporte Vectorial de Mínimos Cuadrados con el cual se realiza el estudio, en las secciones 6.5, 6.6 y 6.7, se exhibe el Clasificador de Soporte Vectorial aplicado a las comunas de Santiago, Puente Alto y La Florida respectivamente, mientras que en 6.8 se exponen las comparaciones de los resultados obtenidos.

Finalmente, en el capítulo 7 se presentan las conclusiones generadas luego del desarrollo del trabajo y de la obtención de los resultados.

2. Máquinas de Soporte Vectorial (SVM).

Las Máquinas de Soporte Vectorial (del inglés Support Vector Machine y sus siglas SVM) fueron desarrolladas en 1995 por Vladimir Vapnik y están basadas en la teoría de aprendizaje estadístico [3], que a su vez corresponden a la familia de los clasificadores lineales. A diferencia de las Redes Neuronales Artificiales, que utilizan durante la fase de entrenamiento el principio de Minimización del Riesgo Empírico (ERM de sus siglas en inglés, Empirical Risk Minimization), las SVM se basan en el principio de Minimización del Riesgo Estructural (SRM de sus siglas en inglés, Structural Risk Minimization), el que ha mostrado un mejor desempeño que el ERM, ya que las SVMs buscan minimizar la probabilidad de una clasificación errónea sobre nuevos ejemplos, a diferencia del ERM que minimiza el error sobre los datos de entrenamiento [4]. O sea, en palabras simples, lo que persigue esta herramienta es el aprendizaje a partir de los datos de entrada, los que pueden presentar características bastantes dispersas, tal como los datos existentes en el estudio, además éstos son separados en 2 grandes conjuntos (clases). Luego, el aprendizaje se logra mediante la búsqueda de alguna dependencia funcional entre un conjunto de vectores con los datos de entrada y de salida, permitiendo así encontrar un espacio lo más amplio posible con el cual se pueda separar los datos pertenecientes a una clase u otra.

Una SVM es un método de aprendizaje supervisado basado en kernel (funciones núcleo), usados tanto para problemas de clasificación como de regresión. En el caso de la clasificación, las funciones de kernel se utilizan usualmente para transformar los datos de entrada a un espacio de características de dimensión mayor en el cual los datos de entrada se vuelven más separables en comparación con el espacio de entrada original, para luego encontrar el hiperplano que los separe, y maximice el margen *m* entre las clases tal como se puede apreciar en la Figura 2.1. Maximizar el margen *m* es un problema de programación cuadrática (QP) y puede ser resuelto por su problema dual introduciendo multiplicadores de Lagrange. La solución del hiperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados vectores de soporte (SV por sus siglas en inglés, Support Vectors). En un principio a los datos utilizados para hallar el hiperplano de decisión se les llama vectores de entrenamiento o aprendizaje. Algunas de las razones por las que este método ha tenido éxito es que no padece de mínimos locales y el modelo sólo depende de los datos con más información, los cuales son los vectores de soporte.

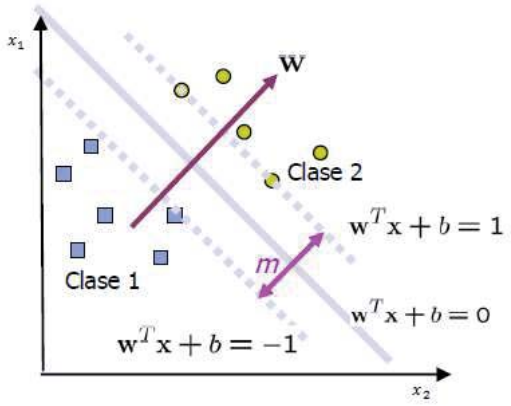

Figura 2.1: Ejemplo de hiperplano de separación entre las clases, en un problema linealmente separable.

Y luego, posterior a la fase de aprendizaje o entrenamiento se comprueba el error cometido tomando otra muestra de datos (denominados conjunto de test o validación) y se compara la salida que se obtiene con la clase original.

Las SVMs han sido desarrolladas como una técnica robusta para clasificación aplicada a grandes conjuntos de datos complejos con ruido; es decir, con variables inherentes al modelo que para otras técnicas aumentan la posibilidad de error en los resultados, pues resulta difícil poder cuantificarlas y observarlas.

Además de sus sólidos fundamentos matemáticos en la teoría de aprendizaje estadístico, las SVMs han demostrado un rendimiento altamente competitivo en un amplio número de aplicaciones de la vida real, tales como bioinformática, minería de texto, reconocimiento facial y procesamiento de imágenes, lo que ha establecido las SVMs como una de las herramientas de última generación en máquinas de aprendizaje y minería de datos, junto con otras técnicas tales como Redes Neuronales y Sistemas Difusos [5]. Aunque cabe destacar que existen aplicaciones en las que las SVMs han demostrado tener mejor desempeño que las técnicas tradicionales como las Redes Neuronales [6] y han sido introducidas como herramientas poderosas para resolver problemas de clasificación. Además las SVM se diferencian de las otras técnicas anteriormente mencionadas ya que no son afectadas por el problema de los mínimos locales, debido a que su entrenamiento se basa en problemas de optimización convexa.

Algunas de las fortalezas de las SVMs son [7]:

- El entrenamiento es relativamente fácil.
- No hay óptimo local como en las redes neuronales.
- Se escalan relativamente bien para datos en espacios dimensionales altos.
- El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente.
- Datos no tradicionales como cadenas de caracteres y árboles pueden ser usados como entrada a la SVM, en vez de vectores de características.

Y dentro de las debilidades se encuentra que, se necesita una buena función kernel, es decir, se necesitan metodologías eficientes para sintonizar los parámetros de inicialización de la Máquina de Soporte Vectorial. Por lo mencionado anteriormente, para este caso los parámetros de la SVM serán estimados mediante la utilización de algoritmos genéticos tales como, Optimización por Enjambre de Partículas (PSO) y algunas de sus variaciones que serán presentadas en capítulos posteriores.

2.1 Máquinas de Soporte Vectorial en Problemas de Clasificación.

Cada vez se hace más común el enfrentarse a problemas en los que es necesario clasificar variados tipos de datos, tal como los que se obtienen en el reconocimiento de voz, diagnóstico médico, procesamiento de imágenes, entre otros. Sin embargo, se ha

demostrado en muchas ocasiones que resolver estos problemas presenta bastantes dificultades [8]. Para hacer frente a estas dificultades es que aparecen las SVM, puesto que se encuentran dentro de las técnicas de clasificación más destacadas y que han tenido gran éxito al ser aplicadas a la resolución de estos problemas. Para clasificación, las SVM permiten obtener clasificadores lineales y no lineales. A continuación se analizarán los casos para datos linealmente separables, linealmente no separables, y posteriormente el clasificador no lineal, donde aparece el concepto de kernel.

2.1.1 Caso Linealmente Separable

Se consideran un número *S* de puntos etiquetados para entrenamiento como se muestra en la Figura 2.2.

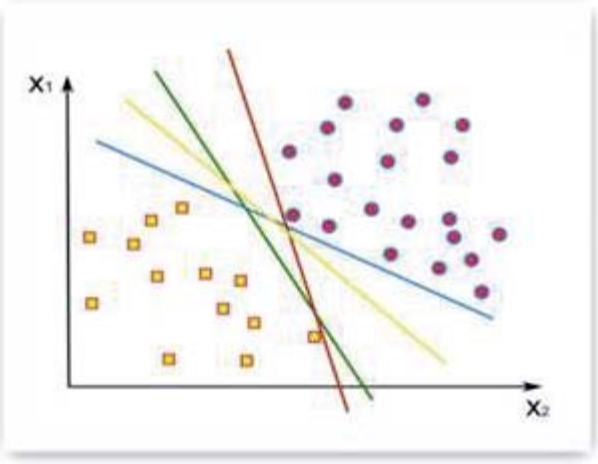

Figura 2.2: Caso linealmente separable.

Cada punto de entrenamiento $x_i \in \Re^N$ pertenece a una de las dos clases y se le ha asignado una etiqueta $y_i \in \{-1,1\}$ para $i = 1, \dots, l$. En la mayoría de los casos, la búsqueda de un hiperplano adecuado en un espacio de entrada es demasiado restrictivo para ser de uso práctico. Una solución a esta situación es mapear el espacio de entrada en un espacio de características de dimensión mayor y ahí buscar el hiperplano óptimo. Sea $z = \varphi(x)$ la notación del correspondiente vector en el espacio de características con un mapeo $\varphi\ de\ \Re^N$ a un espacio de características Z. Se desea encontrar el siguiente hiperplano:

$$f(x) = w \cdot z + b \tag{2.1.1}$$

El que está definido por el par (w, b), tal que sea posible separar el punto x_i de acuerdo a la función:

$$f(x_i) = sign(w \cdot z_i + b) = \begin{cases} 1 & y_i = 1 \\ -1 & y_i = -1 \end{cases} \tag{2.1.2}$$

Donde w ϵ Z y b ϵ ℜ. Más precisamente, el conjunto S se dice que es linealmente separable si existe (w, b) tal que las siguientes inecuaciones sean validas:

$$\begin{cases} (w \cdot z_i + b) \geq 1, & y_i = 1 \\ (w \cdot z_i + b) \leq -1, & y_i = -1 \end{cases} \quad i = 1, \ldots, l \tag{2.1.3}$$

Esto se debe cumplir para todos los elementos del conjunto S. Para el caso linealmente separable de S, es posible encontrar un único hiperplano óptimo, para el cual, el margen entre las proyecciones de los puntos de entrenamiento de dos diferentes clases es maximizado.

En la Figura 2.3 que aparece a continuación se puede apreciar que a partir del hiperplano separador definido por w y b se definen 2 hiperplanos paralelos a éste de tal forma que en los puntos más cercanos al hiperplano x_1 y x_2 se cumpla la siguiente condición $|\langle w, x_i \rangle + b| = 1$. A partir de esto se obtiene que $\langle w, (x_1 - x_2) \rangle = 2$ entonces:

$$\langle \frac{w}{||w||}, (x_1 - x_2) \rangle = \frac{2}{||w||} \tag{2.1.4}$$

Entonces, como se muestra en la Figura 2.3, el margen para esta forma canónica es $\frac{1}{||w||}$. Donde w define el margen de separación óptima y b es el sesgo. La distancia entre el hiperplano de separación y el dato de entrenamiento más cercano al hiperplano es llamado margen.

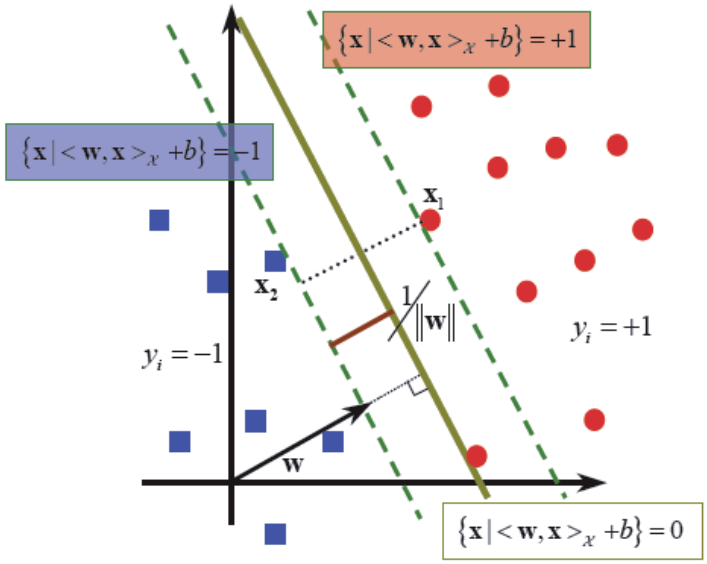

Figura 2.3: Hiperplano canónico y definición del margen geométrico

Sin embargo, pueden existir diferentes hiperplanos que son solución del problema, donde claramente unos son mejores que otros de acuerdo a su margen, donde a mayor margen el hiperplano es mejor. Y tal como fue mencionado anteriormente, es posible encontrar un único hiperplano óptimo llamado hiperplano de separación óptima [9].

Cuando la distancia entre x_1 y x_2 es maximizada, algunos puntos de datos pueden estar sobre x_1 y otros sobre x_2. Estos puntos de datos son llamados vectores de soporte [6] [9], ya que participan de forma directa en definir el hiperplano de separación, los otros puntos pueden ser removidos o cambiados sin cruzar los planos x_1 y x_2 y esto no modificara de alguna forma la habilidad de generalización del clasificador, por lo tanto, la solución de una SVM está dada únicamente por este reducido conjunto de vectores de soporte.

2.1.2 Caso Linealmente No Separable.

En un trabajo de clasificación práctico que los datos sean separables es una situación ideal que no se da normalmente. Por lo tanto es necesario encontrar otra manera de poder resolver estos problemas, la cual se presenta a continuación.

Si el conjunto *S* no es linealmente separable como se muestra en la Figura 2.4, transgresiones a la clasificación deben ser permitidas en la formulación de la SVM.

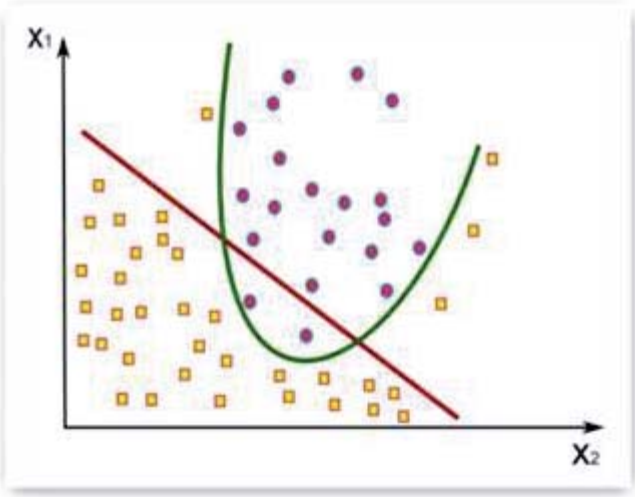

Figura 2.4: Caso linealmente no separable.

Entonces para tratar con datos que no son linealmente separables, el análisis previo puede ser generalizado introduciendo algunas variables no negativas tal como $\xi \geq 0$ de manera que (2.1.3) queda de la siguiente manera:

$$y_i(w \cdot z_i + b) \geq 1 - \xi_i \, , i = 1, \ldots, l \qquad (2.1.5)$$

Los $\xi_i \neq 0$ en (2.1.5) son aquellos para los cuales el punto x_i no satisface (2.1.4). Entonces el término $\sum_{i=1}^{l} \xi_i$ puede ser tomado como algún tipo de medida del error en la clasificación. Entonces el problema de hiperplano óptimo es redefinido de la siguiente manera:

$$\min\left\{\frac{1}{2}w\cdot w + C\sum_{i=1}^{l}\xi_i\right\}$$

$$\text{s.a. } y_i(w\cdot z_i + b) \geq 1-\xi_i,\ i=1,\dots,l \qquad (2.1.6)$$

$$\xi_i \geq 0,\ i=1,\dots,l$$

Donde el parámetro *C* es una constate, y puede ser definido como parámetro de regularización agregado a causa de que en la clasificación para el caso no separable es sumamente probable que se cometan errores, entonces se debe asignar a la función objetivo un costo extra que en cierto modo penalice los errores o el ruido. Este es el único parámetro libre de ser ajustado en la formulación de la SVM. El ajuste de éste puede hacer un balance entre la maximización del margen y la trasgresión a la clasificación [9] [10] . Además este parámetro controla la compensación que ocurre entre la maximización del margen y la minimización del error de entrenamiento. Así, cuando *C* es pequeño, se permite trasgredir el clasificador (clasificar erróneamente) muchas veces, pero a cambio se obtiene un margen grande. Mientras que cuando *C* es grande, no se permite clasificar de manera errónea y se obtiene un margen pequeño. Cabe destacar que *C* es el primer parámetro que debe ser ajustado.

El buscar el hiperplano óptimo en (2.1.6) es un problema de programación cuadrática que puede ser resuelto utilizando multiplicadores de Lagrange (el que se explicará con mayor nivel de detalle más adelante) y transformándolo en el problema dual:

$$Max\ W(\propto) = \sum_{i=1}^{l}\propto_i - \frac{1}{2}\sum_{i=l}^{l}\sum_{j=1}^{l}\propto_i \propto_j y_i y_j z_i\cdot z_j$$

$$\text{s.a. } \sum_{i=1}^{l} y_i \propto_i = 0,\quad 0 \leq \propto_i \leq C,\quad \text{i} = 1,\dots,\text{l} \qquad (2.1.7)$$

Donde $\propto = (\propto_1, \dots \propto_l)$ es un vector de multiplicadores de Lagrange positivos asociados con las constantes en (2.1.5).

2.1.3 Clasificador No Lineal.

Es posible que ocurra en algunas ocasiones que los datos (puntos) no sean linealmente separables en el espacio de entrada. Entonces, cuando esto sucede, existe la posibilidad de transformar los datos a un espacio $\Im$ de características de mayor dimensión, en donde los datos si pueden ser separados por un hiperplano, tal como se muestra en la Figura 2.5

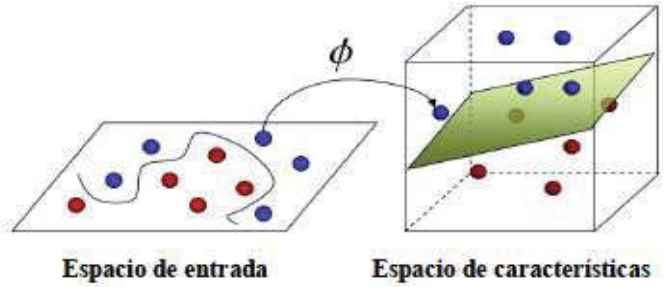

Figura 2.5: Transformación de los datos de entrada a un espacio de mayor dimensión.

Para realizar estas transformaciones se utilizan funciones llamadas funciones núcleo o kernel. Las cuales representan el producto vectorial en el espacio de características. Esta función se define de la siguiente manera:

$$\Phi\colon \Re^D \rightarrow \Im \tag{2.1.8}$$

$$x \rightarrow \Phi(x) \tag{2.1.9}$$

2.2 Tipos de Kernel.

Cuando la SVM es no lineal y los datos no son linealmente separables, la idea principal es llevar los datos de un espacio inicial a otro de mayor dimensión, y esto se logra mediante las funciones núcleo o kernel tal como fue mencionado anteriormente. Según el tipo de kernel y de los valores de sus parámetros se pueden obtener distintas fronteras de decisión. Entre los kernels más utilizados destacan los siguientes:

- Gaussiana:

$$K(x, x') = \exp\left(\frac{-||x-x'||^2}{2\sigma^2}\right) \qquad \sigma > 0 \tag{2.2.1}$$

- Polinomial:

$$K(x, x') = (x^T \cdot x' + c)^d \qquad c \in \Re, \mathrm{d} \in \aleph \tag{2.2.2}$$

- Sigmoidal:

$$K(x, x') = tanh(\mathrm{s}(\mathrm{x}^{\mathrm{T}} \cdot \mathrm{x}') + \mathrm{r}) \qquad s, r \in \Re \tag{2.2.3}$$

- Lineal:

$$K(x, x') = x^T \cdot x \tag{2.2.4}$$

Cabe destacar que cuando se habla de kernel lineal se hace referencia al producto vectorial en el espacio de entrada que equivale a emplear la SVM de margen máximo.

A continuación, en la Figura 2.6 se presenta de manera gráfica como las funciones kernel permiten realizar la separación y el traslado de los datos al espacio de características.

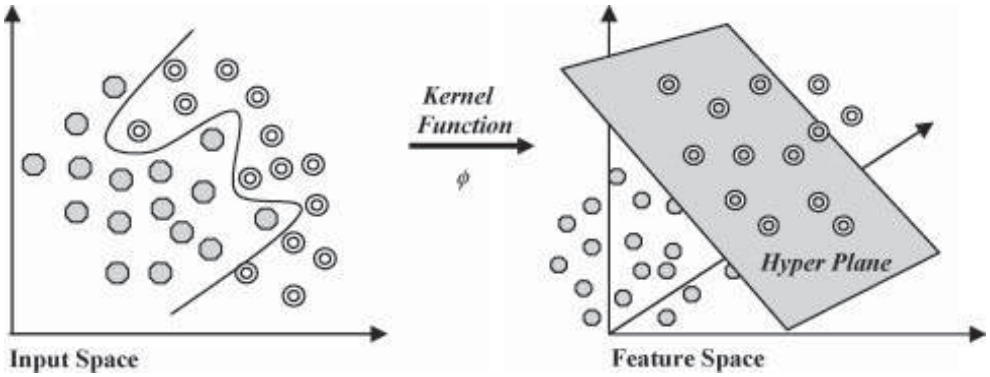

Figura 2.6: Aplicación de la función kernel a los datos.

2.3 Multiplicadores de Lagrange.

Cuando ya se ha definido el problema de optimización que permite la determinación de la superficie óptima para la clasificación de dos clases separables, es necesario encontrar una formulación alternativa del problema, a la que se llama formulación dual. Para encontrar la formulación dual del problema de determinación del clasificador de margen máximo se presenta a continuación la especificación del problema de optimización Lagrangiano, donde se presentaran los conceptos básicos que definen el problema de optimización llamado problema primal de la siguiente forma, $min_x \Phi(x)$ que se encuentra sujeto a, $g_i(x) \geq 0 \quad \forall i \in \{1, \dots, l\}$. Donde $\Phi: \Re^n \to \Re$ es una función objetivo convexa y g_i son restricciones lineales $\Re^n \to \Re$ y, en consecuencia, son convexas, entonces se llama problema de optimización Lagrangiano a:

$$max_\alpha min_x L(\alpha, x) = max_\alpha min_x (\Phi(x) - \sum_{i=1}^{l} \propto_i g_i(x)) \tag{2.3.1}$$

Sujeto a $\propto_i \geq 0 \ \forall i \in \{1, \dots, l\}$, a la función siguiente se le llama Función Lagrangiana:

$$L(\propto, x) = (\Phi(x) - \sum_{i=1}^{l} \propto_i g_i(x)) \tag{2.3.2}$$

Los valores $\propto = (\propto_1, \dots, \propto_l)$ son los denominados Multiplicadores de Lagrange, y $\propto \in R^l$ variable dual.

Entonces este problema de optimización se resuelve mediante los puntos silla de la función de Lagrange, pero como la función objetivo es convexa y las restricciones son lineales (2.3.2) tiene un punto silla único, donde se cumple que $\frac{dL}{dx} = 0$.

Ahora, que ya se han presentado los conceptos básicos, es posible presentar la expresión dual del problema del clasificador de margen máximo. Donde el problema de optimización del clasificador del margen máximo es minimizar:

$$\Phi(w, b) = \frac{||w||^2}{2} \tag{2.3.3}$$

Y está sujeto a:

$$w^t(y_i x_i) \geq 1 + y_i b \quad \forall (x_i y_i) \in S \tag{2.3.4}$$

Luego, es posible reformular las restricciones para que posteriormente permitan construir la función Lagrangiana, entonces con la reformulación queda de la siguiente manera:

$$y_i(w^t x_i + b) - 1 \geq 0 \quad \forall (x_i y_i) \in S \tag{2.3.5}$$

y ahora la función Lagrangiana es:

$$L(w, b, \propto) = \frac{1}{2} w^T w - \sum_{i=1}^{l} \propto_i \{y_i[w^T x_i + b] - 1\} \tag{2.3.6}$$

La búsqueda de un punto de silla óptimo $(w_0, b_0, \propto_0)$ es necesario debido a que el Lagrangiano debe ser minimizado con respecto a w y b, y debe ser maximizado con respecto a $\propto_i$ no negativo, es decir $\propto_i \geq 0$ debe ser encontrado. Este problema se puede resolver, ya sea en un espacio primal, que es el espacio de los parámetros w y b, o en un espacio dual, que es el espacio de los Multiplicadores de Lagrange $\propto_i$. El segundo enfoque entrega resultados profundos, y se considera la solución en un espacio dual. Con el fin de hacer esto, se utilizan las condiciones de Karush Kuhn Tucker (KKT) para el óptimo de la función de restricciones. En este caso entonces, tanto la función objetivo (2.3.6), como la función de restricciones (2.3.5) son convexas, y las condiciones KKT son necesarias y condiciones suficientes para un máximo en (2.3.6). Estas condiciones son, en el punto silla $(w_0, b_0, \propto_0)$, derivadas del Lagrangiano L con respecto a las variables primales, las cuales deberían desaparecer, lo que lleva a:

$$\frac{\partial L}{\partial w_0} = 0 \quad \rightarrow \quad w_0 = \sum_{i=1}^{l} \propto_i y_i x_i \tag{2.3.7}$$

$$\frac{\partial L}{\partial b_0} = 0 \quad \rightarrow \quad \sum_{i=1}^{l} \propto_i y_i = 0 \tag{2.3.8}$$

Y las condiciones complementarias KKT (que indican que en el punto de solución los productos entre variables duales y restricciones son igual a cero) también se deben satisfacer:

$$\propto_i \{y_i[w^T x_i + b] - 1\} = 0 \quad i = 1, \ldots, l \tag{2.3.9}$$

Reemplazando (2.3.7) y (2.3.8) en las variables primales del Lagrangiano $L(w, b, \alpha)$ (2.3.6) se modifican las variables duales del Lagrangiano $L_d(\propto)$,

$$L_d(\propto) = \sum_{i=1}^{l} \propto_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \propto_i \propto_j x_i^T x_j \tag{2.3.10}$$

A fin de encontrar el hiperplano óptimo, el Lagrangiano dual $L_d(\propto)$ debe ser maximizado con respecto a un $\propto_i$ no negativo, y con respecto a las siguientes restricciones:

$$\propto_i \geq 0, \quad i = 1, \ldots, l \tag{2.3.11}$$

$$\sum_{i=1}^{l} \propto_i y_i = 0 \tag{2.3.12}$$

Cabe destacar que el Lagrangiano para el problema de optimización dual $L_d(\propto)$ está expresado en términos de los datos de entrenamiento y depende solamente del producto escalar de los patrones de entrada $(x_i^T x_j)$. Se debe tener en cuenta que el número de variables desconocidas es igual al número de datos de entrenamiento. Y luego, cabe destacar que el número de parámetros libres es igual al número de Vectores de Soporte, pero no depende de la dimensionalidad del espacio de entrada.

2.4 SVM de Mínimos Cuadrados (LS-SVM).

Las SVMs de Mínimos Cuadrados (LS de sus siglas en inglés Least Square) son una modificación de la formulación básica de las SVM y fueron desarrolladas por Suykens y Vandewalle [11] en 1999. Estas se caracterizan porque la optimización lleva a resolver un sistema de ecuaciones lineales más sencillo de utilizar que las soluciones de programación cuadrática normalmente utilizadas en las SVM tradicionales. Además, esta modificación ayuda a resolver problemas asociados a las SVM tradicionales en los cuales el costo de procesamiento era muy alto.

Entonces, en primer lugar, en vez de inecuaciones la formulación de las LS-SVM utiliza ecuaciones de igualdad donde el valor de la derecha de la ecuación se considera como un valor objetivo más que un valor umbral. Sobre este valor objetivo se permite un error de estimación variable ς^k, siendo esta una variable de error. Luego, la formulación del problema de las SVM de Mínimos Cuadrados se presenta de la siguiente manera. Considerando un modelo en el espacio primario:

$$y(x) = w^T \Phi(x) + b \tag{2.4.1}$$

Donde $x \in \Re^n, y \in \Re, \Phi(\cdot): \Re^n \to \Re^{n_h}$ es un mapeo hacia un espacio de características multidimensionales mayor y posiblemente de infinitas dimensiones. Considerando un conjunto de entrenamiento $\{x^k, y^k\}_{k=1}^m$, podemos formular el siguiente problema de optimización en el espacio primario:

$$\min_{w,b,\varsigma} J_p(w,\varsigma) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{l} \varsigma^{k^2} \tag{2.4.2}$$

Tal que: $$y^k = w^T \Phi(x^k) + b + \varsigma^k, \;\; k = 1, \ldots, l$$

Es necesario mencionar que este problema puede no tener solución en el espacio primario, y por lo tanto se pasa al dual mediante la Lagrangiana tal como sigue:

$$L(w,b,\varsigma;\propto) = J_p(w,\varsigma) - \sum_{k=1}^{l} \propto^k w^T \Phi(x^k) + b + \varsigma^k - y^k \tag{2.4.3}$$

Donde $\propto^k$ son los multiplicadores de Lagrange. Y las condiciones de optimización vienen dadas por:

$$\frac{\partial L}{\partial w} = 0 \ \rightarrow w = \sum_{k=1}^{l} \alpha^k \, \Phi(x^k)$$

$$\frac{\partial L}{\partial b} = 0 \ \rightarrow w = \sum_{k=1}^{l} \alpha^k = 0$$

$$\frac{\partial L}{\partial \varsigma^k} = 0 \ \rightarrow \alpha^k = \gamma \varsigma^k, \qquad k = 1, \dots, l$$

$$\frac{\partial L}{\partial \alpha^k} = 0 \ \rightarrow w^T \Phi(x^k) + b + \varsigma * k - y^k = 0, \ \ k = 1, \dots, l \tag{2.4.4}$$

Luego, se obtiene la siguiente solución:

$$\begin{bmatrix} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -Y^T \\ 0 & 0 & C^I & -I \\ Z & Y & I & 0 \end{bmatrix} \begin{bmatrix} w \\ b \\ \varsigma \\ \propto \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vec{1} \end{bmatrix} \tag{2.4.5}$$

Dónde:

$$Z = [\varphi(x_1)^T y_1, \dots, \varphi(x_n)^T y_N], Y = [y_1, \dots, y_N], \vec{1} = [1, \dots, 1], \varsigma = [\varsigma_1, \dots, \varsigma_N],$$
$$\propto = [\propto_1, \dots, \propto_N]$$

La solución está dada también por:

$$\begin{bmatrix} 0 & -Y^T \\ Y & ZZ^T + C^{-1I} \end{bmatrix} \begin{bmatrix} b \\ \propto \end{bmatrix} = \begin{bmatrix} 0 \\ \vec{1} \end{bmatrix} \tag{2.4.6}$$

El modelo LS-SVM para aproximación funcional resultante se convierte entonces en:

$$y(x) = \ \sum_{k=1}^{l} \alpha^k \, K(x, x^k) + b \tag{2.4.7}$$

Donde α^ky b son solución al sistema de ecuaciones lineales (4.4.6).

3. Optimización por Enjambre de Partículas (PSO).

La Optimización por Enjambre de Partículas (PSO de sus siglas en inglés, Particle Swarm Optimization) es una metaheurística evolutiva y de búsqueda [12], que fue desarrollada por Kennedy y Eberhart en 1995 [13]. PSO simula el comportamiento social de organismos presentes en la naturaleza, tal como las bandadas de pájaros, los cardúmenes de peces o los enjambres de abejas, esto con la finalidad de describir un sistema de evolución de forma automática. Cada candidato único a solución, tal como un ave individual de la bandada, es una partícula en el espacio de búsqueda, y cada partícula utiliza su memoria individual y conocimiento adquirido mediante el enjambre en su conjunto para encontrar la mejor solución [14]. Todas las partículas tienen valores fitness (medida de la calidad de la solución), los que son evaluados por una función fitness para ser optimizados, y tienen velocidades que dirigen el movimiento de las partículas en el sentido que corresponda a las mejores soluciones. Durante el movimiento, cada partícula ajusta su posición de acuerdo a su propia experiencia, y al mismo tiempo de acuerdo a la experiencia de las partículas vecinas, haciendo uso de la mejor posición encontrada por sí mismo y por su vecino. Las partículas se mueven a través del espacio del problema siguiendo a las partículas óptimas actuales.

El enjambre inicial es generalmente creado de tal manera que la población de partículas se distribuye aleatoriamente sobre el espacio de búsqueda. En toda iteración, cada partícula es actualizada mediante dos mejores valores, llamados pbest y gbest. Cada partícula realiza un seguimiento de sus coordenadas en el espacio del problema, las que se asocian con la mejor solución (fitness) que la partícula ha alcanzado hasta el momento. Este valor de fitness es almacenado, y es llamado pbest. Cuando la partícula toma la población completa como su vecino topológico, el mejor valor es un valor global y este es llamado gbest. A continuación se presenta el pseudo código del procedimiento PSO.

```
For p = 1 to número de partículas
   If the fitness of x_p es mayor que el fitness de pbest_p
   then Actualizar pbest_p = x_p
   For k ∈ Vecindario de x_p
         If the fitness of x_k es mayor que la de gbest
         then Actualizar gbest = x_k
   Next k
   For each dimensión d
   v_pd^new = w * v_pd^old + c_1 * rand_1 * (pbest_pd − x_pd^old) + c_2 * rand_2 * (gbest_d − x_pd^old)
   if v_pd ∉ (V_min, V_max) then
   v_pd = max (min(V_max, v_pd), V_min)
                                   x_pd = x_pd + v_pd
   Next d
  Next p
Next generar hasta criterio de término
```

Algoritmo 1.1: *Procedimiento PSO.*

Donde, v_{pd}^{new} y v_{pd}^{old} son las velocidades de las partículas, x_{pd}^{old} es la posición actual de la partícula, y x_{pd}^{new}es la nueva posición actualizada de la partícula. Los dos factores $rand_1$

y $rand_2$ son números aleatorios entre (0, 1), mientras que c_1 y c_2 son los factores de aceleración, usualmente $c_1 = c_2 = 2$. Las velocidades de las partículas de cada dimensión son tratados como velocidad máxima V_{max}. Si la suma de las velocidades provoca que el total de la velocidad de esa dimensión exceda V_{max}, entonces la velocidad en esa dimensión se limita a V_{max}, el cual es un parámetro especificado por el usuario.

Los valores funcionales adaptivos están basados en los datos de las características de las partículas que representa la dimensión característica, esta data es clasificada mediante Máquinas de Soporte Vectorial para obtener precisión en la clasificación, la SVM sirve como evaluador de la función fitness de PSO.

Además de la Optimización por Enjambre de Partículas original existen distintas variaciones que buscan mejorar el rendimiento de la optimización realizando distintos cambios al PSO tradicional, estos cambios pueden ser distintas formas de inicializar las partículas o las velocidades, que incluyan nuevos parámetros, o también que tengan distintos métodos de actualización, entre otros.

3.1 Variantes de PSO.

Como era de esperar, el modelo tradicional de PSO ha sufrido modificaciones según distintos autores, los que esperan con estas poder mejorar el rendimiento y los resultados obtenidos realizando distintos cambios ya sea en los parámetros, la forma de inicializar las partículas, o en los distintos métodos de actualización, entre otros. A continuación se presentan las variaciones de PSO con las cuales se va a desarrollar el siguiente trabajo y que serán comparadas además con la versión tradicional de PSO y así poder evaluar los resultados y el desempeño de cada una con las distintas configuraciones.

3.1.1 Quantum PSO (QPSO).

Es una versión de inspiración cuántica (quantum) del algoritmo PSO propuesta relativamente hace no mucho tiempo [15]. El algoritmo QPSO permite a todas las partículas tener un comportamiento cuántico, en lugar de la dinámica clásica que tenía la versión anterior. Así, en lugar del movimiento aleatorio, una especie de movimiento cuántico se aplica en el proceso de búsqueda. Cuando QPSO es probado frente un set de funciones de evaluación comparativa, se ha demostrado un rendimiento superior comparado a la versión clásica de PSO, pero bajo la condición de grandes tamaños de población [15]. Una de las características más atractiva de este nuevo algoritmos es que reduce el número de parámetros de control, entonces estrictamente hablando existe solo un parámetro que debe ser ajustado en QPSO.

3.1.2 Improved PSO (IPSO).

Variante de PSO propuesta por Bin, Zhigang y Xingsheng en [16] propusieron IPSO (Improved Particle Swarm Optimization), donde se plantea un peso de inercia dinámico con el cual se realiza la búsqueda, y que va disminuyendo de acuerdo a como van aumentando las iteraciones. Posee 2 valores de entrada, $w_{inicial}$ y μ, que son definidos desde el principio.

4. Clasificador Vectorial de Soporte (CVS).

Para la resolución del problema planteado en el presente estudio, se utilizarán Máquinas de Soporte Vectorial de mínimos cuadrados (LS-SVM). Esta técnica permite simplificar algunos aspectos de las SVM tradicionales sin perder sus ventajas. Además, para la selección y estimación de los parámetros de la SVM se utilizará PSO, QPSO e IPSO. Por lo tanto, a continuación se presentará el modelo que será la base para construir el Clasificador Vectorial de Soporte propuesto para el desarrollo e implementación del estudio. Luego, se presentarán las métricas que servirán para la evaluación y comparación de los modelos, las que al mismo tiempo servirán para medir el rendimiento obtenido en cada caso.

Los modelos desarrollados que serán presentados en el trabajo son los siguientes: LS-SVM-PSO, LS-SVM-QPSO, LS-SVM-IPSO los que fueron desarrollados mediante la herramienta MATLAB utilizando el framework lssvmlab [17].

4.1 Modelo General.

El modelo que se presenta a continuación tiene como fin optimizar el parámetro γ, que es el encargado de regularizar la máquina vectorial, además de encontrar los parámetros asociados al kernel, tal como fue presentado en secciones anteriores. En este estudio será utilizado principalmente el kernel Gaussiano y como se mencionó al inicio del capítulo para la obtención de los parámetros óptimos del modelo se utilizará PSO. Entonces, se presenta la Figura 4.1 donde se puede ver de manera gráfica como será el modelo LS-SVM con los algoritmos PSO y sus variaciones.

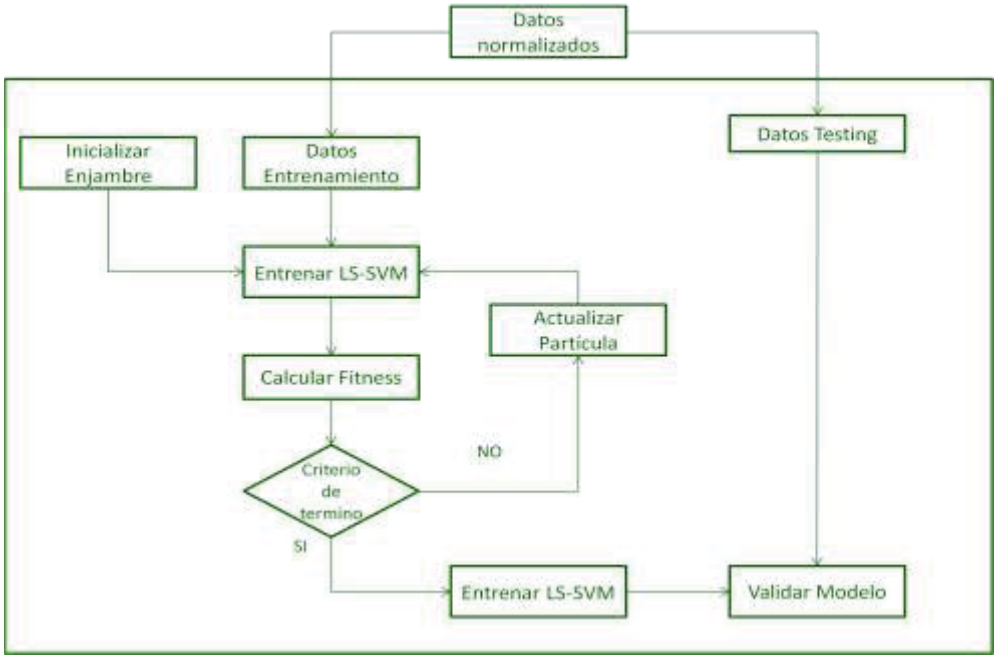

Figura 4.1: Modelo LS-SVM con algoritmos PSO.

4.2 Parámetros de Entrada.

En el caso de PSO, ajustar los parámetros permite optimizar el rendimiento del algoritmo, pero para el primer procesamiento se definen parámetros iniciales que sirven de partida para el modelo, una buena elección de los parámetros C_1 y C_2 puesto que pueden producir una rápida convergencia del algoritmo y evitar mínimos locales, es $C_1 = C_2 = 2$. Para la velocidad, se debe limitar un máximo y mínimo para que los movimientos de las partículas no se acerquen a valores que no presentan utilidad, por lo tanto se define $V_{max} = V_{min} = \pm 1{,}6$. Finalmente, el coeficiente inercial $w = 0{,}8$, donde este a medida que aumentan las iteraciones va disminuyendo lo que provoca un cambio desde un modo de exploración, donde se generan evaluaciones en regiones distantes del espacio de búsqueda, a un modo de explotación en el cual se evalúan soluciones en regiones acotadas y pequeñas con respecto al espacio de búsqueda.

4.3 Métricas de Evaluación.

Para evaluar los distintos modelos en cada una de sus etapas, y comparar los resultados obtenidos es necesario utilizar métricas de exactitud, las que permitirán tener conocimiento del grado de exactitud de la clasificación que se obtenga en los distintos casos. Para esto existen muchos métodos posibles, de los cuales fueron seleccionadas las métricas que se muestran a continuación: Exactitud, Sensibilidad, Especificidad y área de la curva ROC. Estas métricas se forman considerando los siguientes conceptos:

- Verdaderos Positivos (VP): número de éxitos. O sea, corresponde al número de personas que se clasificaron con daños.

- Verdaderos Negativos (VN): número de rechazos correctos. Corresponde a las personas que fueron detectadas ilesas correctamente.

- Falsos Positivos (FP): número de falsos lesionados. En este contexto corresponden al número de personas detectadas como lesionadas, pero que realmente resultaron ilesas.

- Falsos Negativos (FN): número de falsos ilesos. Corresponde al número de personas detectadas como ilesas, pero que realmente resultaron lesionadas.

Los valores mencionados anteriormente, VP, VN, FP y FN se presentan en la siguiente figura, la que se denomina matriz de confusión o tabla de contingencia, donde cada fila de la matriz representa el número de resultados obtenidos en cada clase, mientras que cada columna representa las instancias en la clase real.

		Resultado Real	
		POSITIVO	NEGATIVO
Resultado Obtenido	POSITIVO	**VP**	**FP**
	NEGATIVO	**FN**	**VN**

Figura 4.2: Matriz de Confusión.

A partir de los Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN) se construyen los siguientes ratios, los cuales corresponden a las siguientes métricas:

- **Exactitud:** corresponde al total de personas bien clasificadas, ya sea con lesión o sin lesión, dentro del total de personas. Es representada de la siguiente forma:

$$Exactitud = \left(\frac{VP+VN}{VP+VN+FN+FP}\right) \quad (4.2.1)$$

- **Sensibilidad:** corresponde a la probabilidad de que una persona realmente lesionada sea detectada como tal por la prueba. Es representada por la siguiente ecuación:

$$Sensibilidad = \left(\frac{VP}{VP+FN}\right) \quad (4.2.2)$$

- **Especificidad:** corresponde a la probabilidad de que una persona ilesa sea detectada como tal por la prueba. Es representada por la siguiente ecuación:

$$Especificidad = \left(\frac{VN}{VN+FP}\right) \quad (4.2.3)$$

- **Valor Predictivo Positivo (VPP) o Precisión:** corresponde a la probabilidad de padecer la lesión si se obtiene un resultado positivo en el test. Es representada por la siguiente ecuación:

$$VPP = \left(\frac{VP}{VP+FP}\right) \quad (4.2.4)$$

- **Valor Predictivo Negativo:** corresponde a la probabilidad de que una persona con un resultado negativo en la prueba esté realmente ileso. Es representada por la siguiente ecuación:

$$VPN = \left(\frac{VN}{VN+FN}\right) \quad (4.2.5)$$

- **Fracción de Falsos Positivos** (FFP o "1-Especificidad")**:**

$$FFP = \left(\frac{FP}{FP+VN}\right) \; o \; 1 - Especificidad \tag{4.2.6}$$

Otra de las métricas más usadas en clasificación binaria corresponde a la curva ROC (*Receiver Operating Characteristic)*. La curva ROC es un gráfico en el que se observan todos los pares sensibilidad/especificidad resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados. Se define por FFP y FVP como ejes x e y respectivamente. Representa los intercambios entre verdaderos positivos y falsos positivos. Dado que FVP es equivalente a la "sensibilidad" y FFP es "1-especificidad", el gráfico ROC se llama a veces la representación de (1-Especificidad) frente a la Sensibilidad. Cada resultado de la clasificación de una instancia de la matriz de confusión representa un punto en el espacio ROC. Cada punto de la curva representa un par S/1-E correspondiente a un nivel de decisión determinado. Una prueba con discriminación perfecta, sin solapamiento de resultados en las dos poblaciones, tiene una curva ROC que pasa por la esquina superior izquierda, donde Sensibilidad y Especificidad toman valores máximos (igual a 1).

El Área Bajo la Curva (UAC): A partir de la curva ROC se deriva el "Área Bajo la Curva" (UAC). El área bajo la curva ROC es siempre mayor o igual a 0,5. El rango de valores se mueve entre 1 (discriminación perfecta) y 0,5 (no hay diferencias en la distribución de los valores de la prueba entre los 2 grupos). Se presenta una especie de guía para poder interpretar la curva ROC mediante la interpretación de los siguientes intervalos del UAC.

[0.5, 6)] = Test malo.
[0.6, 0.75] = Test regular.
[0.75, 0.9] = Test bueno.
[0.9, 0.97] = Test muy bueno.
[0.97, 1)] = Test excelente.

Cualitativamente, cuanto más próxima es una curva ROC a la esquina superior izquierda, más alta es la exactitud global de la prueba. De la misma forma, si se dibujan en un mismo gráfico las curvas obtenidas con distintas pruebas, aquella que esté situada más hacia arriba y hacia la izquierda tiene mayor exactitud. De este modo, por simple observación se obtiene una comparación cualitativa. A continuación se presenta un ejemplo de la curva ROC en la Figura 4.3.

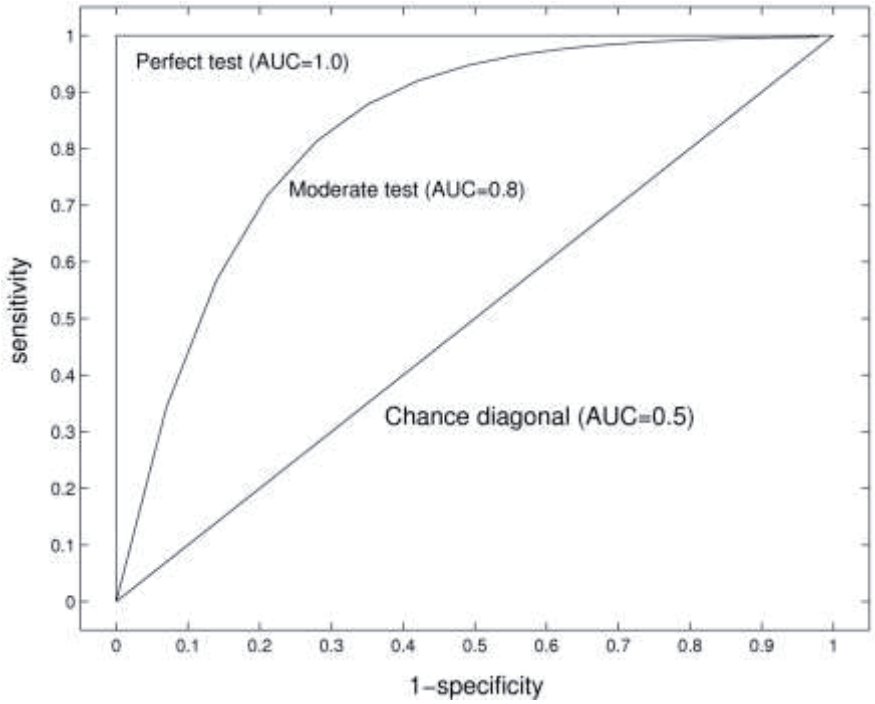

Figura 4.3: Curva ROC.

5. Representación y Explicación de los datos utilizados.

A continuación, se describen los datos de los accidentes de tránsito que fueron considerados para el desarrollo del estudio. Estos datos, en primera instancia se obtienen a partir del registro que genera Carabineros de Chile luego de concurrir al lugar donde ocurren los accidentes de tránsito. La obtención de esta información se realiza de forma manual en planillas que se encuentran predefinidas. Posterior al registro de los datos, estos se entregan a la CONASET, entidad encargada de almacenarlos en sus bases de datos para generar estadísticas y estudios que permitan obtener información que sea de utilidad.

En esta oportunidad, los datos fueron entregados por la Escuela de Ingeniería de Transporte de la Pontificia Universidad Católica de Valparaíso, quienes trabajan directamente con la CONASET y es por esto que tienen acceso a esta información. Los datos se encontraban en planillas Excel, en las que se podían encontrar datos desde el 2003 al 2009, los cuales a su vez se encuentran divididos en 3 sub planillas, estas son Accidentes, Personas y Vehículos, las cuales son explicadas a continuación.

- Accidentes: Aquí se presenta toda la información de las características del accidente y los factores que estuvieron presentes. Los distintos atributos de esta tabla son: identificador del accidente (el cual es un número único para cada accidente), fecha y hora, comuna, si el sector era urbano o rural, la ubicación relativa al momento del accidente, la característica, tipo, estado y la condición de la calzada, el estado atmosférico, la causa del accidente, el tipo de accidente, y el resultado luego del accidente, o sea, si los involucrados resultaron muertos, graves, menos graves, leves o ilesos.
- Personas: En esta tabla se presentan los datos estrictamente relevantes de cada individuo involucrado en el accidente, estas características son: el identificador del accidente (necesario para enlazar a las personas con el accidente en el que estuvieron involucrados), sexo, edad, calidad del involucrado, y el resultado de la misma forma que la tabla accidentes.
- Vehículo: Esta tabla considera la información de el o los vehículos que estuvieron involucrados en los distintos accidentes, los datos aquí presentes son: el identificador del accidente (con el mismo fin que en la tabla persona), tipo de vehículo, el servicio, las consecuencias, dirección en la que viajaba, y la maniobra realizada.

5.1 Estructura de los Datos.

Se detalla a continuación la estructura que tienen los datos y la base de datos original presentada por la CONASET, y la figura con la explicación gráfica de las 3 entidades presentadas anteriormente con los principales atributos que las definen. Además como se puede apreciar en la Figura 5.1, este modelo se puede interpretar de la siguiente manera. En un accidente pueden haber 1 o más personas involucradas y a su vez pueden haber 1 o más vehículos involucrados. A continuación se presentan los atributos de cada entidad y su detalle para mayor entendimiento de los datos.

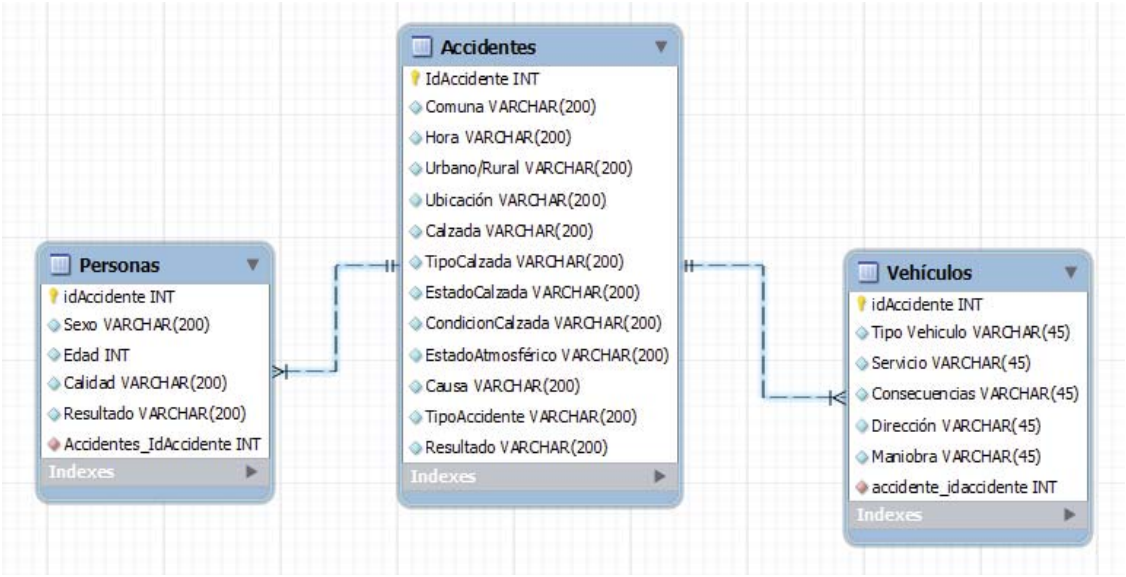

Figura 5.1: Entidades y atributos base de datos CONASET.

- Accidentes. Esta es la entidad principal y corresponde a los accidentes ocurridos. Los atributos que posee son:

  - IdAccidente: es un número con el que se identifican todos los accidentes ocurridos, el cual es único para cada accidente.
  - Fecha: fecha del accidente.
  - Hora: hora a la que se registró el evento.
  - Región: región donde ocurre.
  - Comuna: comuna de la región donde se protagonizó el accidente.
  - Tipo de accidente: dentro de este atributo existen distintos tipos, los que se muestran a continuación:

    - Atropello.
    - Caída.
    - Colisión: la que puede ser: colisión frontal, lateral, por alcance y perpendicular.
    - Impacto con animal.
    - Choque con objeto: el que puede ser: choque con objeto frontal, lateral y posterior.
    - Choque con vehículo detenido: donde se encuentran: choque con vehículo detenido frente/frente, frente/lado, frente/posterior, lado/frente, lado/lado, lado/posterior, posterior/frente, posterior/lado y posterior/posterior.
    - Volcadura.
    - Incendio.
    - Descarrilamiento

- Otro tipo.

- Causa: corresponde a la razón por la que se ocasionó el accidente, y las causas pueden ser las siguientes:

    - Fallas mecánicas: las que pueden ser, fallas en frenos, dirección, falla eléctrica, suspensión, neumáticos, motor y carrocería.
    - Adelantamiento: donde se pueden encontrar adelantamientos sin el espacio o tiempo suficiente, sin efectuar la señal respectiva, por la berma, sobrepasando línea continua, en cruce, curva, cuesta, puente.
    - Conducción: pueden ser bajo la influencia del alcohol, bajo la influencia de drogas o estupefacientes, contra el sentido del tránsito, en estado de ebriedad, condiciones físicas deficientes (cansancio, sueño), por izquierda eje calzada, no atento condiciones del tránsito momento, sin mantener distancia razonable ni prudente y cambiar sorpresivamente pista de circulación.
    - No respetar derecho a paso: aquí se encuentran los derechos a paso al peatón y vehículo.
    - Pasajero: en este caso puede ser si sube o desciende de vehículo en movimiento, viaja en pisadera del vehículo, por imprudencia y ebriedad del pasajero.
    - Peatón: donde pueden ser distintos casos, el peatón permanece sobre la calzada, cruza la calzada de forma sorpresiva o descuidada, imprudencia, ebriedad, cruza la calzada fuera del paso de peatones, y cruza el camino o la carretera sin precaución.
    - Señalización: donde se encuentran, señalización mal instalada o mantenida en forma defectuosa, desobedecer la luz roja de semáforo, desobedecer la indicación de carabinero en servicio, desobedecer la señal ceda el paso, desobedecer la señal pare, desobedecer otra, semáforo en mal estado o deficiente y desobedecer la luz intermitente del semáforo.
    - Velocidad: dentro de ellas se encuentran: mayor que máxima permitida, no razonable ni prudente, no reducir cruce de calles, cumbre, curva, etc., exceso en zona restringida y menor que mínima establecida,
    - Carga: las que pueden ser, mayor que la autorizada para el vehículo, obstruye la visual del conductor, escurre a la calzada, sobresale estructura del vehículo.
    - Virajes indebidos.
    - Animales sueltos en vía pública.
    - Vehículos en retroceso, conducir.
    - Vehículos en panne sin señalización o deficiente.
    - Pérdida control vehículo.
    - Suicidio.
    - Otras causas.
    - Fuga por hecho delictual.

- Ubicación relativa: Se refiere al lugar donde ocurrió el accidente y las distintas ubicaciones pueden ser las siguientes:

    - Cruce con semáforo funcionando.
    - Cruce sin señalización.
    - Enlace a desnivel.
    - Plaza de peaje.
    - Acera o berma.
    - Enlace a nivel.
    - Tramo de vía curva vertical.
    - Túnel.
    - Cruce con señal "pare".
    - Cruce regulado por carabinero.
    - Tramo de vía recta.
    - Acceso no habilitado.
    - Tramo de vía curva horizontal.
    - Rotonda.
    - Otros no considerados.
    - Cruce con señal "ceda el paso".
    - Cruce con semáforo apagado.
    - Puente.

- Estado Atmosférico: el estado atmosférico del día en el momento del accidente, los cuales pueden ser:

    - Despejado.
    - Nublado.
    - Lluvia.
    - Llovizna.
    - Neblina.
    - Nieve.

- Tipo calzada: Existen seis tipos de calzada, estos son:

    - Concreto.
    - Asfalto.
    - Adoquín.
    - Mixto.
    - Ripio.
    - Tierra.

- Estado Calzada: estado de cómo está la calzada al momento del accidente, estos pueden ser:

    - Bueno.
    - Regular.
    - Malo.

- Condición Calzada: las distintas condiciones en la que se encontraba la calzada al momento de ocurrido el accidente, y estas pueden ser:

        - Seco.
        - Húmedo.
        - Mojado.
        - Con barro.
        - Con nieve.
        - Con aceite.
        - Escarcha.
        - Gravilla.
        - Otros.

    - Muertos: cantidad de personas muertas resultado del accidente.
    - Graves: cantidad de personas graves resultado del accidente.
    - Menos Graves: cantidad de personas con lesiones menos graves resultado del accidente.
    - Leves: cantidad de personas leves resultado del accidente.
    - Ilesos: cantidad de personas ilesas que se vieron involucradas en el accidente.

- Personas: Esta entidad se refiere a las personas que están involucradas en un accidente en específico. Se pueden diferenciar los distintos atributos entre las personas:

    - Calidad: se refiere a la calidad o condición que tenían las personas involucradas en el accidente, estas son:

        - Peatón.
        - Conductor.
        - Pasajero.

    - Sexo.

        - Masculino.
        - Femenino.

    - Resultado: corresponde a los distintos estados en los que puede quedar una persona luego de ocurrido el accidente, estos son:

        - Muertos.
        - Graves.
        - Menos Graves.
        - Leves.
        - Ilesos.

- Vehículos: Entidad asociada a los vehículos involucrados en un accidente en particular. Un accidente por lo menos presenta uno o más vehículos. Los atributos y características asociadas a esta entidad son:

- Tipo vehículo: se refiere al tipo del vehículo o los vehículos que se vieron involucrados en el accidente, estos pueden ser:

    - Bus/ taxi bus, minibús, trolebús, automóvil, camioneta, jeep, furgón, ambulancia, camión simple, camión simple con remolque, tracto-camión, tracto-camión con remolque, carro bomba, carro transporte de valores, remolque/semi remolque, motocicleta, motoneta/bicimoto, moto arenera, bicicleta, tracción animal, carro tracción humana, tractor, maquinaria agrícola, maquinaria movimiento tierras, maquinaria industrial, patín/patineta, patín motorizado, ferrocarril, dado a la fuga, otros no clasificados.

- Servicio: son los distintos tipos de servicios que cumplían los vehículos involucrados en los accidentes, los cuales pueden ser:

    - Carabineros, fiscal, particular, transporte escolar, taxi básico, taxi colectivo urbano, taxi colectivo rural, bomberos, salud, locomoción colectivo urbano, locomoción colectivo rural, servicio interurbano, servicio internacional, carga normal, carga peligrosa, dado a la fuga, otros sin especificar.

6. Desarrollo del Clasificador de Soporte Vectorial (CVS).

En este capítulo se presenta el proceso de desarrollo de los modelos realizados en el estudio, donde se considera la preparación correspondiente de los datos, puesto que no venían en el formato necesario para ser introducidos en el Clasificador de Soporte Vectorial. Por lo tanto en primera instancia fue necesario realizar una limpieza de la data, para luego codificarla tal como se muestra más adelante. Posteriormente se presenta la función de Validación Cruzada, la cual fue la función fitness utilizada para evaluar los modelos que fueron desarrollados.

6.1 Preparación de los Datos.

Anterior a la implementación, es necesario realizar un trabajo previo con los datos, o sea, es necesario prepararlos para que estos puedan ser utilizados por la SVM, esta preparación considera dejar los datos con el formato correspondiente para lo que se desea realizar en el trabajo, a este paso se le denomina pre-procesamiento de los datos. En esta etapa se realizaron un número de pasos que serán explicados a continuación. Cabe destacar que la cantidad de datos existentes es aproximadamente de unos 75.000 registros, considerados entre el 2003 y 2009, esto se debe a que la región Metropolitana es la ciudad con más habitantes, por lo tanto con mayor cantidad de vehículos y en consecuencia con mayor cantidad de accidentes del país.

6.1.1 Pre-Procesamiento de los Datos.

Considerando los datos recibidos en las planillas Excel, los que fueron procesados dentro de la misma herramienta utilizando funciones que esta posee. En primer lugar fueron eliminados los datos en los que existían atributos con valores nulos puesto que estos no presentan utilidad para el trabajo a desarrollar. Luego, los datos fueron filtrados por región, puesto que se deben considerar solo los datos correspondientes a las comunas de la región Metropolitana ya que esta es el foco del estudio. Posteriormente fueron eliminados los datos en los que se encontraron anomalías como por ejemplo que la edad de alguna persona excediera el límite lógico, puesto que hubo casos en que la edad era de 999 años, por lo tanto ese tipo de datos fue removido.

6.1.2 Selección de Datos y Codificación.

Tomando en cuenta que el modelo a clasificar corresponde al estado en el que resultan las personas que se vieron involucradas en los accidentes de tránsito, el que puede ser, con daños o ilesos. Posteriormente se deben tomar en cuenta los atributos más relevantes y que

permita representar de la mejor manera el modelo. A raíz de esto serán considerados para el estudio los atributos que se presentan a continuación:

1. Comuna.
2. Urbano/Rural.
3. Estado Atmosférico.
4. Hora.
5. Causa.
6. Tipo Calzada.
7. Estado Calzada.
8. Condición Calzada.
9. Tipo de Accidente.
10. Calidad.
11. Sexo.
12. Edad.
13. Resultado.

Una vez seleccionados solo los atributos a utilizar, se debe tener en cuenta que para el trabajo son necesarios utilizar los datos codificados, puesto que para poder trabajar con la SVM se debe contar con datos cuantitativos mientras que los datos originales son cualitativos. Por lo tanto, a continuación los datos de cada atributo son codificados.

De estos 13 atributos que se consideraran para llevar a cabo el siguiente trabajo, cabe destacar que el número 13, o sea, el atributo Resultado será considerado como nuestro objetivo o clase. Este está compuesto por todas las posibilidades que fueron mostradas con anterioridad para el atributo Resultado, las que pueden ser Muertos, Graves, Menos Graves, Leves e Ilesos. Entonces el atributo Resultado queda de la siguiente manera:

- Resultado:
  - Muertos. = 1
  - Graves. = 1
  - Menos Graves. = 1
  - Leves = 1

  - Ilesos. = -1

Aquí se puede apreciar que todo Resultado con el código 1 será considerado como un solo resultado, puesto que como se dijo en el inicio de esta sección, se quiere clasificar el estado en el que resultan las personas luego del accidente, o sea, con daño o ilesos. Por lo tanto a una "Persona con Daños" se le asigna el código 1 mientras que a una "Persona Ilesa" se le asigna el código -1.

A continuación se presentan unos ejemplos para ver como resultan luego de la codificación los otros atributos. Por ejemplo el atributo Sexo y Calidad que aparecen con sus códigos correspondientes quedan de la siguiente manera:

- Sexo:
    - Masculino = 1
    - Femenino = 2

- Calidad:
    - Conductor = 1
    - Peatón = 2
    - Pasajero = 3

Otro punto que es necesario mencionar, es el ajuste realizado al atributo Hora, puesto que la complejidad de trabajar con la hora como números era muy alta, entonces para facilitar el uso de esta, se separó en los siguientes rangos horarios, Madrugada, Mañana, Tarde y Noche, para los cuales se consideraron los siguientes rangos y códigos respectivamente.

- Madrugada: 0:00:00 – 7:59:59 = 1
- Mañana: 8:00:00 – 11:59:59 = 2
- Tarde: 12:00:00 – 17:59:59 = 3
- Noche: 18:00:00 – 23:59:59 = 4

6.2 Datos a Utilizar.

Dados los antecedentes presentados por CONASET en su sitio web [2] y lo observado en los datos existentes, la mayor cantidad de accidentes de la Región Metropolitana ocurren en las siguientes comunas, Santiago, Puente Alto, La Florida, Maipú, Las Condes, por lo tanto los datos de estas comunas fueron separados y en un principio nos centraremos principalmente en estos. A continuación en la Figura 6.1 se presenta un mapa de la Región Metropolitana donde se pueden apreciar las comunas con la mayor cantidad de accidentes para validar lo anteriormente mencionado.

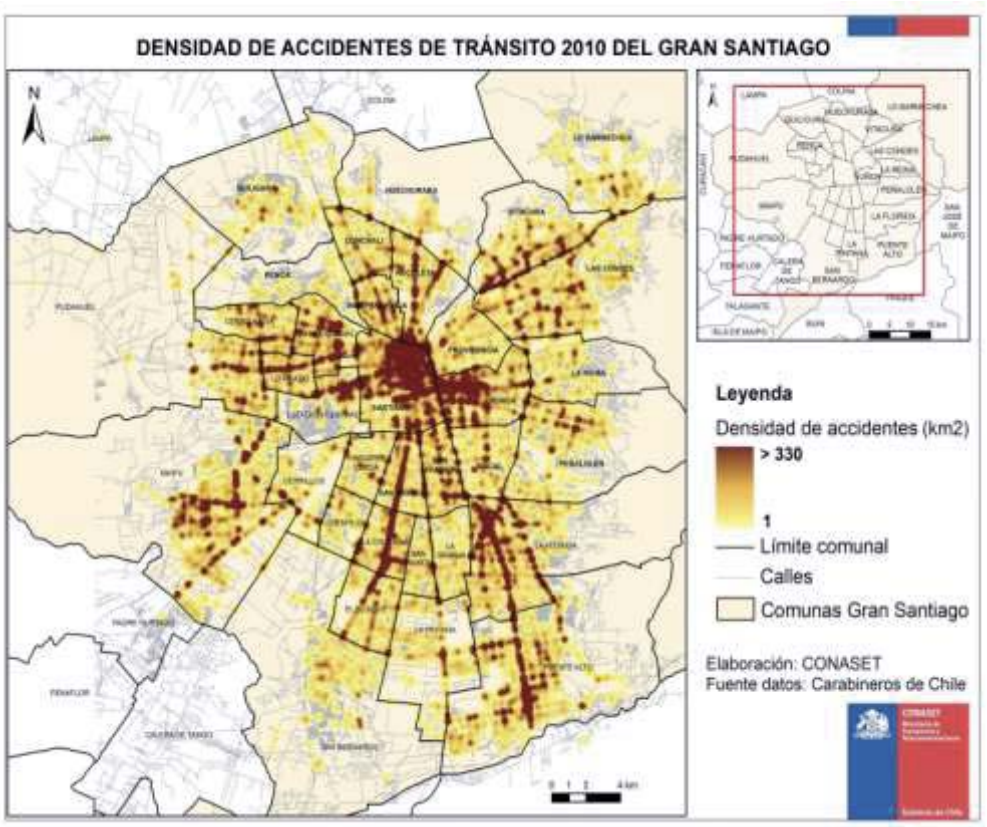

Figura 6.1: Densidad de Accidentes de Tránsito Región Metropolitana.

Y en la Figura 6.2 se presenta donde existe mayor densidad de personas lesionadas en los accidentes de tránsito ocurridos en la Región Metropolitana, y como se puede apreciar a simple vista, ambas imágenes son bastante similares entre sí, puesto que donde existe un mayor número de accidentes es a su vez donde existe la mayor cantidad de personas lesionadas resultado de estos accidentes.

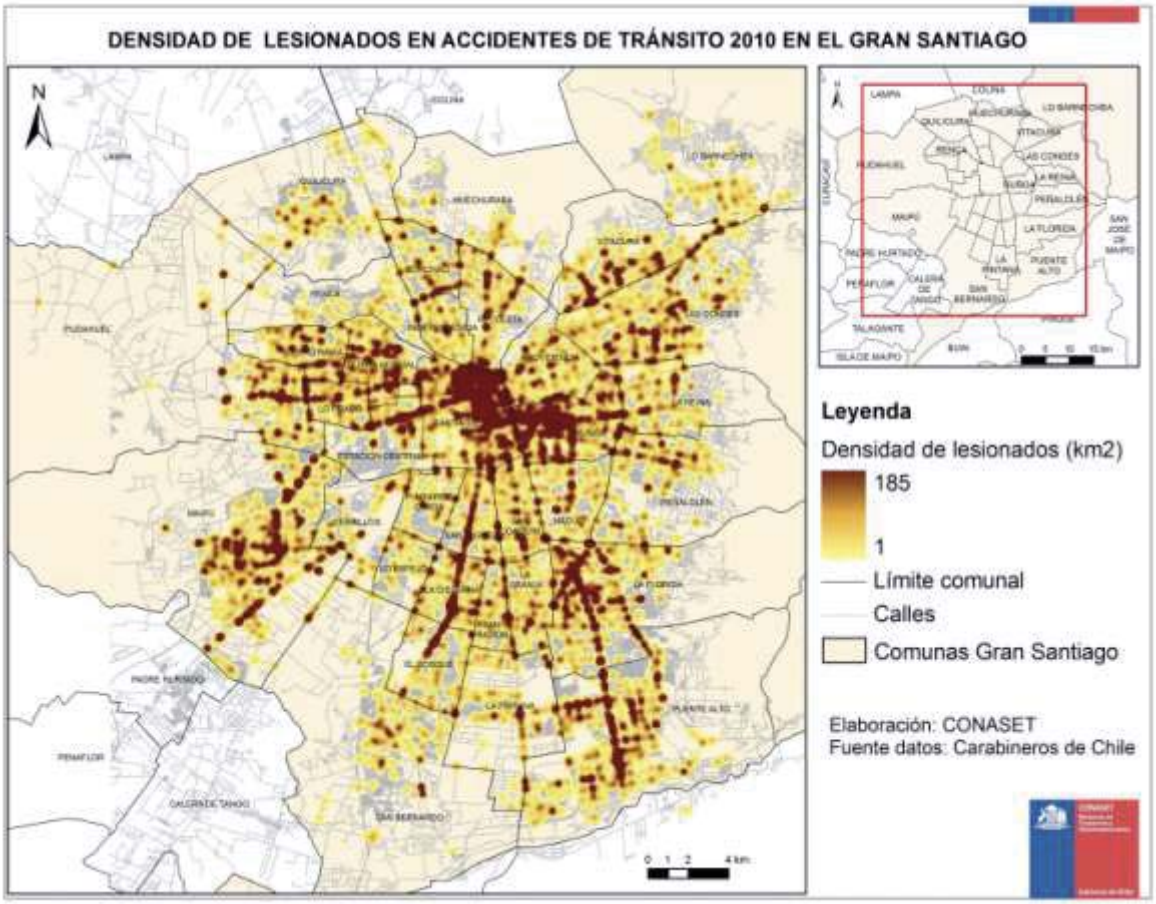

Figura 6.2: Densidad de Lesionados en Accidentes de Tránsito en la Región Metropolitana.

Por lo tanto, considerando lo anterior y lo explicitado en las imágenes nos centraremos en las comunas más representativas puesto que estas prestan mayor información para el trabajo a realizar, estas son Santiago, Puente Alto y la Florida. Además, en la discusión de resultados esto puede ser bastante útil puesto que al segmentar por comuna se pueden tomar mejores decisiones y/o soluciones que se quieran implementar para reducir el número de accidentes y la gravedad de las consecuencias de estos.

6.3 Función de Costo o Fitness.

La función de fitness corresponde a la función que es utilizada para evaluar la calidad de la solución que obtiene el Clasificador Vectorial de Soporte. En este caso en particular se utilizará la Validación Cruzada, función con la cual serán evaluados los 3 modelos que serán presentados a continuación.

La función fitness antes mencionada, consiste básicamente en dividir los datos de entrenamiento en forma aleatoria en L partes. En la $i-ésima(i=1,\dots,l)$ iteración la $i-ésima$ parte de los datos es usada para realizar la validación y las $l-1$ son asignados

como datos de entrenamiento, donde se mide el costo de los elementos mal clasificados. De la misma forma, se itera hasta la última parte del conjunto de datos, rescatando el costo en cada iteración. Finalmente se realiza un promedio de los costos de las distintas iteraciones y este se retorna al algoritmo PSO (o sus variaciones). Cabe destacar que en un principio del estudio se utilizó también otra función de fitness, MSE (error cuadrático medio) pero ésta en ningún caso presentó mejores resultados que la Validación Cruzada, por lo cual se descartó del trabajo final y solo fue considerada esta última.

6.4 Modelo General LS-SVM.

Debido a la necesidad de encontrar los parámetros para cada uno de los 3 modelos realizados fue necesario realizar pruebas en cada uno de los casos. Por lo tanto, durante la etapa de entrenamiento de los tres modelos se realizó un gran número de iteraciones con la finalidad de buscar el conjunto óptimo de parámetros para la etapa posterior de prueba o testing, para lo que se utilizaron combinaciones de los siguientes parámetros que se muestran en la Tabla 6.1.

Parámetros de Entrenamiento	
n° de Partículas	10-20-30
Número iteraciones	80-100-150
% de Entrenamiento	70% a 90%

Tabla 6.1: Parámetros de Entrenamiento.

Con los resultados obtenidos en esta etapa, se logró encontrar los parámetros que serán utilizados en los 3 modelos para el resto del trabajo, y serán presentados más adelante.

Por otro lado, es necesario destacar que luego de las primeras pruebas se determinó que ejecutar el algoritmo con todos los registros correspondientes a 1 año no era posible, puesto que la herramienta y la máquina no eran capaces de realizar este trabajo por la gran cantidad de información que se debía manejar además del manejo de matrices de gran dimensión. Lo anterior generaba problemas de memoria y la ejecución no se podía realizar con éxito Por lo tanto la decisión de considerar solamente las comunas más representativas de la región Metropolitana presentó grandes beneficios puesto que el trabajo se podía realizar con mayor rapidez ya que el número de datos se reduce en gran medida.

6.4.1 Modelo LS-SVM PSO.

En este modelo se utiliza el algoritmo PSO tradicional para la optimización de los parámetros necesarios para ejecutar la máquina LS-SVM, por lo que la partícula del enjambre está compuesta por dos parámetros que se inicializan aleatoriamente, estos son γ y σ^2.

Se realizaron gran cantidad de iteraciones en etapas de training de la máquina vectorial para poder encontrar los parámetros óptimos propios de PSO con los que se realizaría a continuación la etapa de testing. Dentro de todas las ejecuciones se obtuvo que los parámetros óptimos son los que se muestran en la Tabla 6.2. Estos fueron los óptimos obtenidos con los que se realizó el proceso de entrenamiento para la LS-SVM con PSO. Además, se muestra el criterio de término de las ejecuciones, el cual corresponde al número máximo de iteraciones.

Parámetros PSO	
Número de partículas	20
Número iteraciones	100
Coeficiente inercial (w)	0,8
Componente cognitiva (c1)	2
Componente social (c2)	2
Velocidad máxima	1,6
% de entrenamiento	90%

Tabla 6.2: Parámetros PSO.

6.4.2 Modelo LS-SVM QPSO.

La diferencia entre PSO y QPSO está en el algoritmo de búsqueda de los parámetros γ y σ^2, puesto que en QPSO se realiza la optimización basándose solamente en la posición de la partícula, es decir, la velocidad de PSO es dejada de lado. Pero ambos modelos se asemejan, ya que utilizan la misma configuración de sus partículas, es decir, se componen de γ y σ^2, inicializadas aleatoriamente. Y a continuación se muestran los parámetros óptimos con los que se ejecutarán las pruebas para la LS-SVM QPSO.

Parámetros QPSO	
Número de partículas	20
Número iteraciones	100
Alfa 1	0,6
Alfa 2	1
% de entrenamiento	90%

Tabla 6.3: Parámetros QPSO.

6.4.3 Modelo LS-SVM IPSO.

IPSO posee 2 valores de entrada, los que fueron inicializados con los siguientes valores, $w_{inicial} = 0{,}8\ y\ \mu = 1{,}0012$. Además, tiene por característica un peso de inercia dinámico, es decir, el peso de inercia va disminuyendo conforme se incrementa la generación iterativa. También se debe considerar que un valor alto del peso de inercia privilegia la

exploración global y un peso de inercia más pequeño privilegia la exploración local. En IPSO, todas las partículas comparten información mutuamente a nivel global y se favorece con descubrimientos y experiencias previas de las partículas aledañas durante el proceso de búsqueda. Esta variante de PSO también utiliza los parámetros γ y σ^2, los cuales se inicializan aleatoriamente. Se muestran en la siguiente tabla los parámetros para este algoritmo.

Parámetros IPSO	
Número de partículas	20
Número iteraciones	100
Coeficiente inercial (w)	0.8
μ	1.00012
Componente cognitiva (c1)	2
Componente social (c2)	2
Velocidad máxima	1,6
% de entrenamiento	90%

Tabla 6.4: Parámetros IPSO.

6.5 Modelos Aplicados a la Comuna de Santiago.

6.5.1 LS-SVM PSO para Santiago.

Para buscar el mejor número de partículas se realizaron 10 ejecuciones, variando los números de partículas para cada caso entre 10, 20 y 30, pero se presentará solo el LS-SVM PSO para la comuna de Santiago, puesto que para los otros modelos y comunas se comportó de manera similar.

w=0,8	iteraciones=100	partículas=10		
Run	**Error Mínimo**	**Error Máximo**	**Desviación Estándar**	**Error Promedio**
1	0,21	0,27	0,016971753	0,2322
2	0,24	0,28	0,011806521	0,254
3	0,21	0,26	0,035355339	0,26
4	0,22	0,31	0,025469134	0,2509
5	0,24	0,33	0,028056761	0,2637
6	0,21	0,25	0,028284271	0,2262
7	0,23	0,3	0,019668721	0,2501
8	0,21	0,24	0,008187204	0,2242
9	0,23	0,27	0,012065059	0,2467
10	0,22	0,3	0,021387242	0,2454

Tabla 6.5: Ejecuciones Respecto del Error con 10 Partículas.

w=0,8	iteraciones=100	partículas=20		
Run	**Error Mínimo**	**Error Máximo**	**Desviación Estándar**	**Error Promedio**
1	0,23	0,29	0,01651782	0,2533
2	0,23	0,29	0,016058919	0,2413
3	0,22	0,24	0,014142136	0,24
4	0,22	0,27	0,014664945	0,2397
5	0,22	0,3	0,021948286	0,2447
6	0,22	0,26	0,028284271	0,229
7	0,21	0,25	0,010760947	0,2256
8	0,24	0,27	0,008528029	0,244
9	0,21	0,26	0,015311909	0,2267
10	0,21	0,26	0,016775102	0,2329

Tabla 6.6: Ejecuciones Respecto del Error con 20 Partículas.

w=0,8	iteraciones=100	partículas=30		
Run	**Error Mínimo**	**Error Máximo**	**Desviación Estándar**	**Error Promedio**
1	0,22	0,25	0,007282884	0,2343
2	0,22	0,27	0,015123732	0,2366
3	0,23	0,29	0,042426407	0,29
4	0,25	0,28	0,005659532	0,2523
5	0,22	0,31	0,025783872	0,2528
6	0,22	0,29	0,049497475	0,2361
7	0,22	0,27	0,013666667	0,2347
8	0,24	0,28	0,012427974	0,2497
9	0,24	0,24	3,90536E-16	0,24
10	0,23	0,28	0,01431394	0,2454

Tabla 6.7: Ejecuciones Respecto del Error con 30 Partículas.

En la Figura 6.3 se puede apreciar cómo se comporta el error promedio en la etapa de entrenamiento a medida que el número de partículas va en aumento. Con esto se pudo determinar que el número de partículas óptimo es 20. Del mismo modo, los otros 2 modelos se comportaron de manera similar.

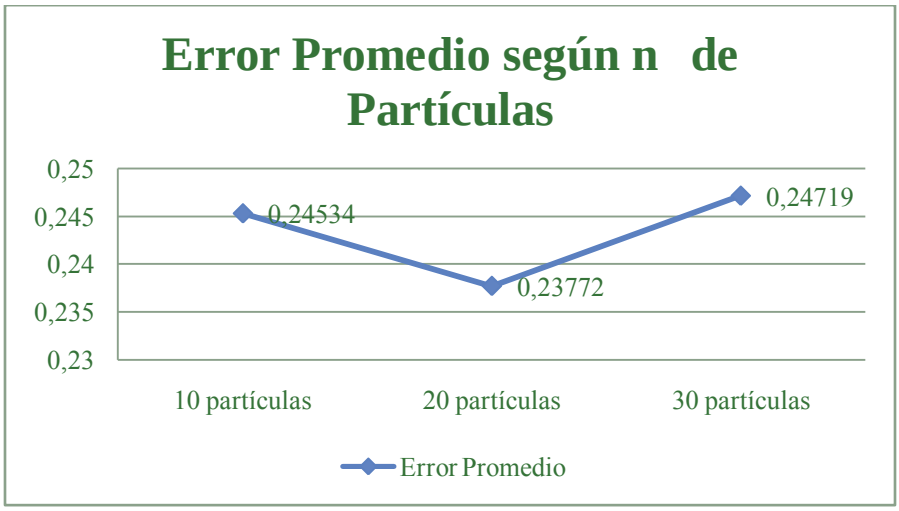

Figura 6.3: Error Promedio Según N° de Partículas.

En la siguiente Tabla 6.8 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,84	1104,8	193,76	48	5	68	17	22	0,7	0,1	0,86	775,51
0,81	733,36	169,32	46	4	66	22	26	0,68	0,06	0,86	761,64
0,81	167,61	147,87	44	7	68	19	26	0,70	0,09	0,85	759,4
0,81	1292,82	140,37	43	10	69	16	26	0,73	0,13	0,87	761,92
0,80	661,54	220,94	50	5	61	22	27	0,69	0,08	0,85	764
0,80	812,96	153,59	57	4	54	23	27	0,71	0,07	0,88	801,61
0,80	321,99	190,63	46	1	64	27	28	0,63	0,02	0,84	761,8
0,79	297,36	119,28	74	15	143	44	59	0,63	0,09	0,82	510,75
0,79	55,23	147,57	40	5	69	24	29	0,63	0,07	0,84	764,83
0,79	1418,07	183,18	43	7	66	22	29	0,66	0,10	0,82	764,73

Tabla 6.8: Resumen Mejores Resultados PSO para Santiago.

El mejor resultado obtenido con una exactitud de 84% es presentado a continuación con su correspondiente matriz de confusión (Según lo presentado en la Figura 4.2).

	POSITIVO	NEGATIVO	Total		
POSITIVO	48	5	53	**PRECISIÓN**	0,906
NEGATIVO	17	68	85	**VPN**	0,80
Total	65	73	138		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	0,74	0,93			

Tabla 6.9: Matriz de Confusión del Mejor Resultado PSO para Santiago.

De la Tabla 6.9 se obtiene que el clasificador entrega un resultado bastante aceptable, puesto que se han clasificando de manera correcta el 84% de las personas que resultan en el estado lesionado o ileso luego de haberse visto afectados por una accidente, por otra parte la sensibilidad nos muestra que un 74% de personas fueron bien clasificadas y que resultaron lesionadas tras ocurrido el accidente. A continuación se presenta la especificidad, ésta resultó ser un 93% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. Luego, relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 90,6% de estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de 80%, lo que significa que de las personas detectadas como ilesas el 80% de estas están realmente ilesas. También se puede apreciar que los mejores valores obtenidos para los parámetros γ y σ^2 son 1104,8 y 193,76 y el tiempo de ejecución es de 13 minutos aproximadamente.

6.5.2 LS-SVM QPSO para Santiago.

En la siguiente Tabla 6.10 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,82	889,86	36,24	55	10	58	15	25	0,79	0,15	0,89	765,28
0,80	944,4	43,68	50	8	60	20	28	0,71	0,12	0,83	772,86
0,79	618,4	40,27	43	17	66	12	29	0,78	0,20	0,85	772,06
0,79	722,94	59,047	40	14	69	15	29	0,73	0,17	0,81	772,14
0,78	604,08	26,36	46	12	61	19	31	0,71	0,16	0,80	780,91
0,78	391,87	13,33	51	11	57	19	30	0,73	0,16	0,85	812,28
0,77	812,19	50,09	52	15	54	17	32	0,75	0,22	0,80	785,22
0,76	813,60	32,05	51	14	54	19	33	0,73	0,21	0,80	766,65
0,76	484,25	49,66	47	17	58	16	33	0,75	0,23	0,82	767,65
0,76	1274,74	213,29	44	11	61	22	33	0,67	0,15	0,83	766,60

Tabla 6.10: Resumen Mejores Resultados QPSO para Santiago.

El mejor resultado obtenido con una exactitud de 82% es presentado a continuación con su correspondiente matriz de confusión. Además para este valor se obtuvo que los mejores parámetros obtenidos fueron γ=944,4 y σ^2=43,68 con un tiempo de ejecución de 13 minutos.

	POSITIVO	NEGATIVO	Total		
POSITIVO	55	10	65	**PRECISIÓN**	0,846
NEGATIVO	15	58	73	**VPN**	0,79
Total	70	68	138		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	0,79	0,85			

Tabla 6.11: Matriz de Confusión del Mejor Resultado QPSO para Santiago.

De la Tabla 6.11 se obtiene que el clasificador entrega un buen resultado, se han clasificado correctamente el 82% de las personas que resultan en el estado lesionado o ileso luego de participar en un accidente. Por otra parte presenta un muy buen resultado en la sensibilidad, el cual es de un 79%, lo que indica que esta cifra corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas tras ocurrido el accidente de tránsito. La especificidad resultó ser de un 67% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso, la razón de esto se debe al bajo número de registros de personas ilesas, menos de la mitad de los lesionados. Luego, relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 84,6% de estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de un 79%, lo que significa que de las personas detectadas como ilesas el 79% resulta realmente ilesas después de ocurrido el accidente.

6.5.3 LS-SVM IPSO para Santiago.

En la siguiente Tabla 6.12 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,83	440,26	37,11	49	9	65	15	24	0,77	0,12	0,88	884,10
0,81	540,79	28,66	49	11	63	15	26	0,77	0,15	0,87	769,55
0,79	863,85	21,15	43	11	66	18	29	0,70	0,14	0,78	765,89
0,78	10634	33,37	49	14	59	16	30	0,75	0,19	0,82	762,74
0,78	857,52	91,66	40	8	68	22	30	0,65	0,11	0,79	763,36
0,78	756,74	4,57	58	20	49	11	31	0,84	0,29	0,84	767,81
0,77	589,34	26,84	44	5	62	27	32	0,62	0,07	0,82	763,56
0,77	952,02	35,36	44	10	62	22	32	0,67	0,14	0,83	765,53
0,77	578,49	59,58	52	11	54	21	32	0,71	0,17	0,82	763,39
0,76	68,12	64,72	39	15	66	18	33	0,68	0,19	0,82	763,34

Tabla 6.12: Resumen Mejores Resultados IPSO para Santiago.

El mejor resultado obtenido con una exactitud de 83% es presentado a continuación con su correspondiente matriz de confusión. Además los mejores valores para los parámetros γ y σ^2 fueron 440,26 y 37,11 respectivamente y se demoró aproximadamente 15 minutos.

	POSITIVO	NEGATIVO	Total		
POSITIVO	49	9	58	**PRECISIÓN**	0,845
NEGATIVO	15	65	80	**VPN**	0,81
Total	64	74	138		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	0,77	0,88			

Tabla 6.13: Matriz de Confusión del Mejor Resultado IPSO para Santiago.

De la Tabla 6.13 se obtiene que el clasificador entrega un buen resultado, se han clasificado correctamente el 83% de las personas que resultan en el estado lesionado o ileso luego de participar en un accidente. Por otra parte presenta un muy buen resultado en la sensibilidad, el cual es de un 77%, lo que indica que esta cifra corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas tras ocurrido el accidente de tránsito. La especificidad resultó ser de un 88% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso, la razón de esto se debe al bajo número de registros de personas ilesas, menos de la mitad de los lesionados. Luego, relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 84,5% de estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de un 81%, lo que significa que de las personas detectadas como ilesas el 81% resulta realmente ilesas después de ocurrido el accidente.

6.6 Modelos Aplicados a la Comuna de Puente Alto.

6.6.1 LS-SVM PSO para Puente Alto.

En la siguiente Tabla 6.14 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,89	38,36	169,56	74	9	25	3	12	0,96	0,26	0,89	407,47
0,87	831,89	163,04	70	11	27	3	14	0,96	0,29	0,85	416,90
0,86	1346,78	121,67	64	15	31	1	16	0,98	0,33	0,86	401,28
0,85	270,88	121,84	69	15	25	2	17	0,97	0,38	0,83	412,79
0,85	937,18	120,31	75	13	19	4	17	0,95	0,41	0,78	387,90
0,83	434,01	126,47	71	15	21	4	19	0,95	0,42	0,79	417,01
0,83	734,37	118,56	71	14	21	5	19	0,93	0,40	0,80	412,78
0,83	1429,26	162,44	66	17	26	2	19	0,97	0,40	0,83	396,35
0,83	166,36	172,23	67	17	25	2	19	0,97	0,40	0,86	417,16
0,83	611,54	154,6	57	15	35	4	19	0,93	0,30	0,86	386,79

Tabla 6.14: Resumen Mejores Resultados PSO para Puente Alto.

El mejor resultado obtenido con una exactitud de 89,18% es presentado a continuación con su correspondiente matriz de confusión.

	POSITIVO	NEGATIVO	Total		
POSITIVO	74	9	83	**PRECISIÓN**	0,892
NEGATIVO	3	25	28	**VPN**	0,89
Total	77	34	111		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	0,96	0,74			

Tabla 6.15: Matriz de Confusión del Mejor Resultado PSO para Puente Alto.

De la Tabla 6.15 se obtiene que el clasificador entrega un buen resultado, se han clasificando correctamente el 89,18% de las personas que resultan en el estado lesionado o ileso luego de haberse visto involucrados en un accidente, por otra parte la sensibilidad nos muestra con este 96% que a esta cifra corresponde el porcentaje de personas bien clasificadas y que resultaron lesionadas tras ocurrido el accidente de tránsito. Si bien la especificidad no resultó ser tan alto igual se mantiene dentro de un valor aceptable, y esta resultó ser un 74% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso, la razón de esto se debe al bajo número de registros de personas que resultan ilesas, menos de la mitad de los que resultan lesionados. Luego, relativo a la seguridad del resultado, la precisión del 89,2% indica que del porcentaje de personas detectadas como lesionadas el 89,2% de estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de 89%, lo que significa que de las personas detectadas como ilesas el 90% de estas están realmente ilesas. Para este modelo se puede apreciar también que los mejores valores obtenidos para los parámetros γ y σ^2 son 38,36 y 169,56 respectivamente, y el tiempo aproximado de ejecución fueron 7 minutos. En la Figura 6.4 se presenta la Curva ROC para este que fue el mejor resultado obtenido, y como se fue apreciar el resultado obtenido es de un 89,04%, valor que está muy cerca de valor 1 en la esquina superior izquierda del grafico, el cual es el óptimo ideal. Además existe un 89,04% de probabilidades de que las personas realmente resulten lesionadas, por lo tanto se puede considerar que se está realizando una buena clasificación.

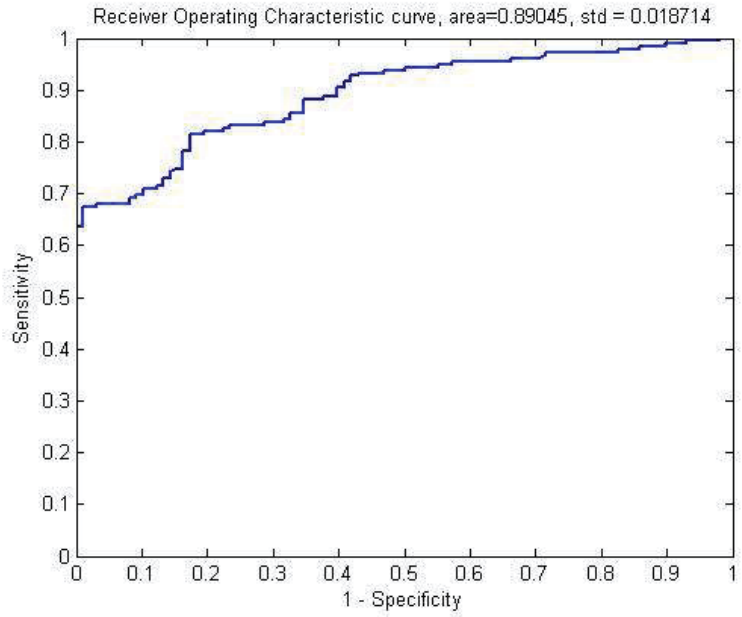

Figura 6.4: Curva ROC para el mejor resultado PSO para Puente Alto.

6.6.2 LS-SVM QPSO para Puente Alto.

En la siguiente Tabla 6.16 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,94	641,27	51,51	70	7	34	0	7	1,00	0,17	0,91	416,41
0,88	2255,46	52,35	66	11	32	2	13	0,97	0,26	0,91	395,06
0,86	632,16	127,03	74	13	21	3	16	0,96	0,38	0,82	418,13
0,86	493,06	63,11	68	12	27	4	16	0,94	0,31	0,82	396,87
0,86	987,78	48,39	71	14	24	2	16	0,97	0,37	0,88	395,73
0,86	873,33	105,51	69	14	27	1	15	0,99	0,34	0,88	399,36
0,85	432,3	76,7	64	12	30	5	17	0,93	0,29	0,87	413,98
0,82	624,64	148,27	64	19	27	1	20	0,98	0,41	0,84	384,81
0,82	494.29	142,56	67	17	24	3	20	0,96	0.41	0,78	403,93
0,82	1001.06	53,6	75	16	16	4	20	0,95	0.50	0,83	385,07

Tabla 6.16: Resumen Mejores Resultados QPSO para Puente Alto.

El mejor resultado obtenido con una exactitud de 94% es presentado a continuación con su correspondiente matriz de confusión.

	POSITIVO	NEGATIVO	Total		
POSITIVO	70	7	77	**PRECISIÓN**	0,909
NEGATIVO	0	34	34	**VPN**	1,00
Total	70	41	111		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	100	0,83			

Tabla 6.17: Matriz de Confusión del Mejor Resultado QPSO para Puente Alto.

De la Tabla 6.17 se obtiene que el clasificador entrega un resultado bastante bueno puesto que se han clasificado correctamente el 94% de las personas que resultan en el estado lesionado o ileso luego de participar en un accidente. Por otra parte presenta un resultado excelente para la sensibilidad 100%, lo que indica que todas las personas fueron bien clasificadas y estas resultaron lesionadas tras ocurrido el accidente de tránsito. La especificidad no resultó ser tan alta como la sensibilidad, pero igualmente se mantiene dentro de un valor aceptable, y esta resultó ser un 83% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso, la razón de esto se debe al bajo número de registros de personas ilesas, menos de la mitad de los lesionados. Luego, relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 90,9% de estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de un 100%, lo que significa que de las personas detectadas como ilesas el 100% resulta realmente ilesas después de ocurrido el accidente. Los mejores valores para los parámetros γ y σ^2 fueron 641,27 y 51,51 respectivamente. El tiempo que demoró esta ejecución fue de 7 minutos aproximadamente.

6.6.3 LS-SVM IPSO para Puente Alto.

En la siguiente Tabla 6.18 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,88	482,92	98,34	76	9	22	4	13	0,95	0,29	0,82	412,61
0,86	444,67	99,87	72	13	23	3	16	0,96	0,36	0,82	381,63
0,86	609,41	80,59	70	11	26	4	15	0,95	0,3	0,82	408,19
0,85	1269,53	83,64	70	11	24	6	17	0,92	0,31	0,81	378,33
0,85	717,19	42,19	67	15	27	2	17	0,97	0,36	0,86	390,3
0,84	1282,13	61,2	70	12	23	6	18	0,92	0,34	0,83	396,26
0,84	1173,42	56,93	74	16	19	2	18	0,97	0,46	0,88	383,17
0,83	1462,82	41,86	61	17	31	2	19	0,97	0,35	0,9	402,65
0,82	830,29	56,76	68	16	23	4	20	0,94	0,41	0,86	391,69
0,82	1350,44	59,78	69	19	22	1	20	0,99	0,46	0,8	394,26

Tabla 6.18: Resumen Mejores Resultados IPSO para Puente Alto.

El mejor resultado obtenido con una exactitud de 88% es presentado a continuación con su correspondiente matriz de confusión. Además los mejores valores para los parámetros γ y σ^2 fueron 482,92 y 98,34 respectivamente y se demoró aproximadamente 7 minutos.

	POSITIVO	NEGATIVO	Total		
POSITIVO	76	9	85	**PRECISIÓN**	0,894
NEGATIVO	4	22	26	**VPN**	0,85
Total	80	31	111		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	0,95	0,71			

Tabla 6.19: Matriz de Confusión del Mejor Resultado IPSO para Puente Alto.

De los resultados en la Tabla 6.19 se puede aprecia que el clasificador obtuvo buenos resultados, considerando también que la exactitud es 88%, lo que quiere decir que el modelo ha clasificado de manera correcta el 88% de las personas que resultan en el estado lesionado o ileso luego de haber participado en un accidente. Por otra parte, la sensibilidad obtenida en esta caso fue 95%, y esta es la cifra que corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas tras ocurrido el accidente. A continuación se presenta la especificidad, y esta resultó ser un 71% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. Luego, relativo a la seguridad del resultado, se tiene un valor predictivo positivo o precisión de 89,4%, y esto significa que de este porcentaje de personas detectadas como lesionadas, estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de 85%, lo que significa que de las personas detectadas como ilesas el 85% de estas están realmente ilesas.

6.7 Modelos Aplicados a la Comuna de La Florida.

6.7.1 LS-SVM PSO para La Florida.

En la siguiente Tabla 6.20 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,89	718,79	137,87	77	10	31	4	14	0,95	0,24	0,92	570,93
0,85	171,48	198,89	73	13	31	5	18	0,94	0,30	0,88	855,06
0,84	333,87	222,1	64	19	38	1	20	0,98	0,33	0,84	497,18
0,84	1465,56	192,12	77	16	25	4	20	0,95	0,39	0,76	500,08
0,83	717,23	134,8	65	16	36	5	21	0,93	0,31	0,81	497,53
0,83	1273,99	200,39	65	21	36	0	21	1,00	0,37	0,87	496,11
0,83	225,89	127,07	68	13	33	8	21	0,89	0,28	0,88	500,03
0,83	611,54	154,6	57	15	35	4	19	0,93	0,30	0,86	386,79
0,81	1134,62	154,94	70	12	29	11	23	0,86	0,29	0,81	504,90
0,81	630,48	68,21	64	10	35	13	23	0,83	0,22	0,89	585,41

Tabla 6.20: Resumen Mejores Resultados PSO para La Florida.

El mejor resultado obtenido con una exactitud de 89%, demoró aproximadamente 10 minutos en ejecutarse y los mejores parámetros γ y σ^2 fueron 718,79 y 137, 87 respectivamente. Se presenta a continuación su correspondiente matriz de confusión.

	POSITIVO	NEGATIVO	Total		
POSITIVO	77	10	87	**PRECISIÓN**	0,885
NEGATIVO	4	31	35	**VPN**	0,89
Total	81	41	122		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	0,95	0,76			

Tabla 6.21: Matriz de Confusión del Mejor Resultado PSO para La Florida.

De la Tabla 6.21 se obtiene que el clasificador entrega un resultado bastante aceptable, puesto que se han clasificando de manera correcta el 89% de las personas que resultan en el estado lesionado o ileso luego de haberse visto afectados por una accidente, por otra parte la sensibilidad nos muestra que el 95% corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas tras ocurrido el accidente. A continuación se presenta la especificidad, y esta resultó ser un 76% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. Luego, relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 88,5% de estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de 89%, lo que significa que de las personas detectadas como ilesas el 89% de estas están realmente ilesas.

6.7.2 LS-SVM QPSO para La Florida.

En la siguiente Tabla 6.22 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,84	897,49	47,17	67	12	35	8	20	0,89	0,26	0,92	497,11
0,84	246,65	93,01	62	12	40	8	20	0,89	0,23	0,84	506,1
0,83	650,72	71,14	56	1	45	20	21	0,74	0,02	0,91	493,79
0,80	10518,61	237,07	63	23	35	1	24	0,98	0,40	0,81	500
0,80	387,33	97,09	61	11	37	13	24	0,82	0,23	0,80	691,65
0,80	18938,69	48,02	59	14	38	11	25	0,84	0,27	0,85	724,5
0,79	848,77	225,52	58	4	38	22	26	0,72	0,10	0,93	585,67
0,79	849,46	54,58	58	12	38	14	26	0,81	0,24	0,83	492,32
0,79	858,25	503,23	67	24	29	2	26	0,97	0,45	0,82	534,01
0,79	10850,95	52,46	56	5	40	21	26	0,73	0,11	0,87	492,79

Tabla 6.22: Resumen Mejores Resultados QPSO para La Florida.

El mejor resultado obtenido con una exactitud de 84% es presentado a continuación con su correspondiente matriz de confusión. Además los mejores valores para los parámetros γ y σ^2 fueron 897,49 y 47,17 respectivamente y se demoró aproximadamente 8 minutos.

	POSITIVO	NEGATIVO	Total		
POSITIVO	67	12	79	**PRECISIÓN**	0,848
NEGATIVO	8	35	43	**VPN**	0,81
Total	75	47	122		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	0,89	0,74			

Tabla 6.23: Matriz de Confusión del Mejor Resultado QPSO para La Florida.

De la Tabla 6.23 se obtiene que el clasificador entrega un buen resultado, se han clasificado correctamente el 84% de las personas que resultan en el estado lesionado o ileso luego de participar en un accidente. Por otra parte presenta un muy buen resultado en la sensibilidad, el cual es de un 89%, lo que indica que esta cifra corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas tras ocurrido el accidente de tránsito. La especificidad resultó ser de un 74% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso, la razón de esto se debe al bajo número de registros de personas ilesas, menos de la mitad de los lesionados. Luego, relativo a la seguridad del resultado, la precisión indica que del porcentaje de personas detectadas como lesionadas un 84,8% de estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de un 81%, lo que significa que de las personas detectadas como ilesas el 81% resulta realmente ilesas después de ocurrido el accidente.

6.7.3 LS-SVM IPSO para La Florida.

En la siguiente Tabla 6.24 se muestra un resumen de los mejores resultados obtenidos para el modelo con los mejores parámetros obtenidos en las etapas de entrenamiento.

Exactitud	γ	σ^2	VP	FP	VN	FN	Errores	TPR	FPR	Área ROC	Tiempo (segundos)
0,83	398,36	25,94	60	8	41	13	21	0,82	0,16	0,95	497,86
0,81	1193,84	89,70	62	10	37	13	23	0,83	0,21	0,89	491,18
0,8	368,07	39,57	55	10	42	15	25	0,79	0,19	0,91	487,36
0,78	1453,23	87,76	65	14	30	13	27	0,83	0,32	0,86	502,09
0,78	30,11	92,17	62	14	33	13	27	0,83	0,3	0,8	577,41
0,78	1148,28	46,56	62	11	33	16	27	0,79	0,25	0,89	593,08
0,77	1152,44	73,76	64	13	30	15	28	0,81	0,3	0,83	510
0,75	893,66	71,73	59	16	33	14	30	0,81	0,33	0,85	505,74
0,75	482,82	79,33	65	12	27	18	30	0,78	0,31	0,84	587,83
0,75	1140,55	47,94	62	11	30	19	30	0,77	0,27	0,84	493,28

Tabla 6.24: Resumen Mejores Resultados IPSO para La Florida.

El mejor resultado obtenido con una exactitud de 83% es presentado a continuación con su correspondiente matriz de confusión. Además los mejores valores para los parámetros γ y σ^2 fueron 482,92 y 98,34 respectivamente y se demoró aproximadamente 8 minutos.

	POSITIVO	**NEGATIVO**	**Total**		
POSITIVO	60	8	68	**PRECISIÓN**	0,882
NEGATIVO	13	41	54	**VPN**	0,76
Total	73	49	122		
	SENSIBILIDAD	**ESPECIFICIDAD**			
	0,82	0,84			

Tabla 6.25: Matriz de Confusión del Mejor Resultado IPSO para La Florida.

De los resultados que se muestran en la Tabla 6.25 se puede aprecia que el clasificador obtuvo buenos resultados, considerando también que la exactitud es 83%, lo que quiere decir que el modelo ha clasificado de manera correcta el 83% de las personas que resultan en el estado lesionado o ileso luego de haber participado en un accidente. Por otra parte, la sensibilidad obtenida en esta caso fue 82%, y esta es la cifra que corresponde al porcentaje de personas bien clasificadas y que resultaron lesionadas tras ocurrido el accidente. A continuación se presenta la especificidad, y esta resultó ser un 82% lo que corresponde al porcentaje de personas bien clasificadas en el estado ileso. Luego, relativo a la seguridad del resultado, se tiene un valor predictivo positivo o precisión de 88,2%, y esto significa que de este porcentaje de personas detectadas como lesionadas, estas realmente resultaron lesionadas luego del accidente, y finalmente, el valor predictivo negativo fue de 76%, lo que significa que de las personas detectadas como ilesas el 76% de estas están realmente ilesas.

6.8 Comparación Modelos.

A continuación se presentan gráficos con los cuales se pueden comparar los distintos modelos según las 3 comunas en relación a algunas métricas. Para esto se utilizaron los datos presentados en los puntos anteriores utilizando los datos de las tablas en las que se presentaban los resúmenes con los mejores resultados.

6.8.1 Comparación por Exactitud.

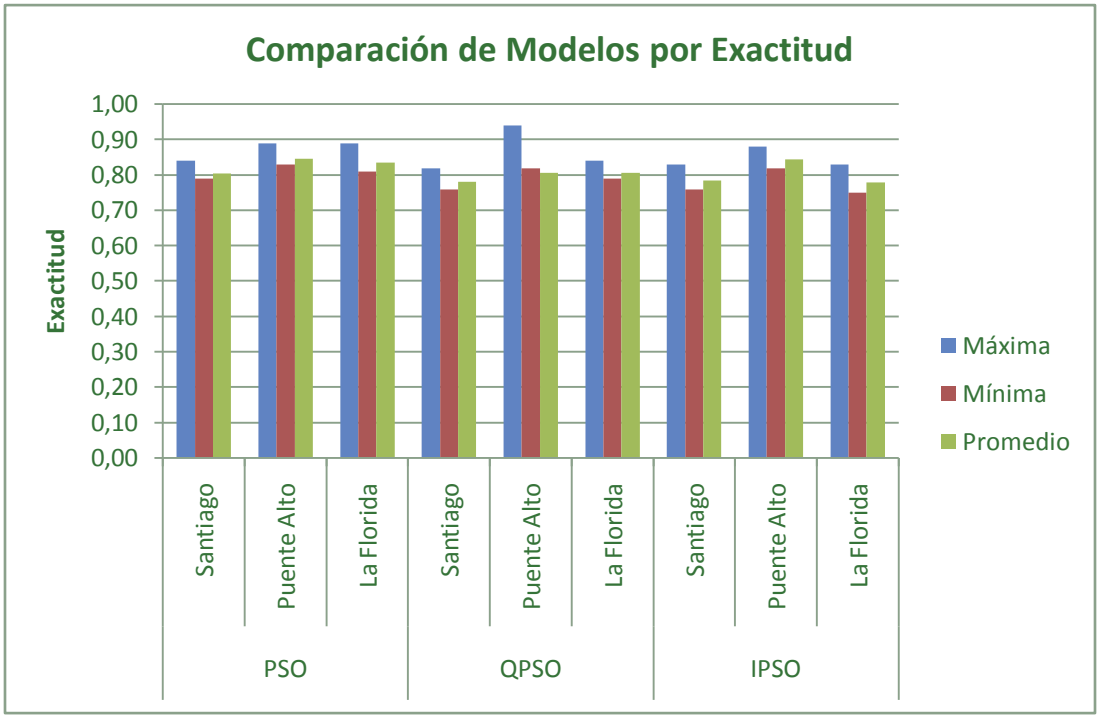

Figura 6.5: Comparación de los Modelos por Exactitud.

De este gráfico se puede decir que los mejores resultados fueron obtenidos con PSO, puesto que presenta los mejores valores en general, los máximos y mínimos. Mientras que el mejor valor individual obtenido fue para el modelo QPSO en la comuna de Puente Alto con un 94%, pero hay que tener en cuenta que para este caso el valor mínimo era el que presentaba la mayor diferencia del máximo, esta diferencia era de 12 puntos porcentuales.

6.8.2 Comparación por Tiempo.

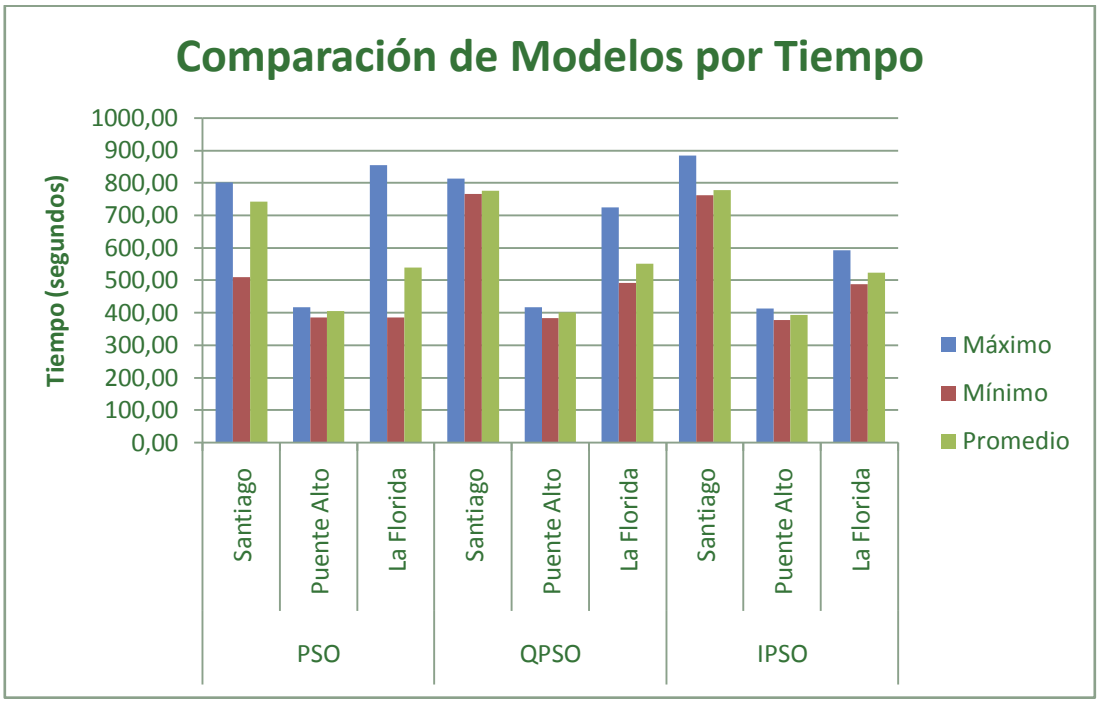

Figura 6.6: Comparación de Modelos por Tiempo.

En el gráfico anterior se presentan los tiempos de ejecución por cada modelo y comuna analizada. Entonces, se puede ver que PSO es el más costoso en relación al tiempo máximo que se presenta, y luego, se puede ver que IPSO es el modelo que en promedio es menos costoso.

7. Conclusiones.

Se ha logrado establecer el estudio del estado del arte en general de los componentes básicos con los que se debe contar para realizar el desarrollo de un modelo basado en Máquinas de Soporte Vectorial (SVM) para la clasificación de accidentes de tránsito, además realizando la optimización de los parámetros de esta SVM mediante algoritmos evolutivos, tal como PSO y algunas de sus variaciones. Luego, se ha desarrollado el modelo del clasificador propuesto, y con esto se da inicio a la correspondiente implementación.

Lo dicho anteriormente relacionado con el desarrollo del modelo se basa puntualmente en el estudio realizado sobre LS-SVM y PSO con sus variaciones, permitiendo la construcción de los modelos que permitieron el desarrollo del trabajo planificado y la obtención de resultados, que permitieran realizar una comparación para la obtención del mejor clasificador dentro de los distintos modelos que fueron generados.

Los resultados obtenidos luego de evaluar los modelos, se encuentran en un buen nivel considerando las 3 alternativas que fueron consideradas, presentando leves variaciones entre sí. El mejor resultado puntual fue presentado por el modelo LS-SVM QPSO, mientras que el modelo LS-SVM PSO presentó los mejores resultados en promedio para las 3 comunas.

Con estos resultados que fueron presentados, se puede apreciar que las Máquinas de Soporte Vectorial permiten obtener buenos resultados para realizar clasificación de datos, pero si no fuera por la estimación de los parámetros mediante la utilización de PSO, y sus variaciones, los resultados obtenidos no hubiesen sido tan alentadores. Por lo que se puede concluir que los algoritmos evolutivos, particularmente los que fueron utilizados para este trabajo, presentan gran utilidad y alto desempeño al momento de ser utilizados en problemas de optimización, con lo que se demuestran las potencialidades de los modelos desarrollados, ya que con la obtención de buenos parámetros se lograrán buenos resultados.

Por lo tanto, el mejor resultado obtenido fue para el modelo LS-SVM QPSO en la comuna de Puente Alto con un 94% de exactitud, 100% de sensibilidad y 83% de especificidad, con lo que se puede concluir que las Máquinas de Soporte Vectorial, utilizando algoritmos evolutivos tal como PSO para la obtención de los parámetros, presentan gran utilidad para la clasificación de datos.

Comparado con otros sistemas clasificadores, las SVM son muy eficientes desde diversas perspectivas. El proceso de aprendizaje es un proceso matemático definido que permite obtener el mejor clasificador, no tan solo un buen clasificador como se obtiene en muchos entrenamientos de redes neuronales, claramente hay que tener presente que esta mejor solución es dependiente del kernel utilizado y de los parámetros que sean escogidos, los cuales pueden variar dependiendo de la experiencia de quién realice el estudio y también de los resultados obtenidos en pruebas previas. Por otra parte, una vez obtenido el modelo, es muy fácil implementarlo en diferentes sistemas. Además se debe destacar que el tiempo de entrenamiento es relativamente corto y que posee una alta velocidad de ejecución en la clasificación de grandes conjuntos de datos.

Este tipo de estudio podría ser de bastante utilidad, en primer lugar a nivel estadístico puesto que permitiría contrastarlo con los datos que existen en la actualidad, y en segundo lugar aún más importante, para que las autoridades tomen decisiones sobre cambios o mejoras que realizar en las calles del país con las cuales se puedan disminuir los accidentes o la gravedad de los mismos. Además de realizar mayor control en las vías en las que estos accidentes se presentan en mayor cantidad. Esto se puede realizar en un principio a nivel de comunas, y posteriormente en sectores más específicos a medida que sean detectados estos focos con mayor cantidad de accidentes.

Finalmente, cabe mencionar que existen variadas opciones para desarrollar nuevos proyectos relacionados a este tema, entre los cuales destacan, la utilización de datos actualizados que permitan validar los resultados obtenidos y además considerar nuevos puntos que puedan presentar mayor cantidad de accidentes. También, se podría realizar un análisis de los datos iniciales antes de ser filtrados, de manera que se pueda determinar qué características permiten que la clasificación sea más eficiente, con lo que se tendría mayor claridad en las causas que determinan la ocurrencia de los accidentes. Otra alternativa relacionada a este tema, sería utilizar otro tipo de herramientas para las distintas fases del desarrollo del trabajo, como por ejemplo para la obtención de los parámetros, con las que se puedan obtener nuevos resultados que permitan realizar comparaciones y así poder generar aún mejores clasificadores. O también, realizar un cambio más drástico y variar la herramienta de fondo, y utilizar por ejemplo Redes Neuronales que permitan tener modelos paralelos con los cuales se pueda conseguir una comparación general del resultado de los distintos clasificadores.

8. Bibliografía

[1] Organización Mundial de la Salud. Los Accidentes de Tránsito Entre las Principales Causas de Murte Entre los Jóvenes. Disponible vía web en http://www.who.int/mediacentre/news/releases/2009/adolescent_mortality_20090911/es/index.html, 2009

[2] Comisión Nacional de Seguridad de Tránsito. Estadísticas Generales, Cantidad de siniestros de tránsito y de víctimas. Diposble vía web en http://www.conaset.cl/portal/portal/default/estadisticas_generales, 2000 - 2010.

[3] Vapnik., V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[4] Vapnik, V. The Nature of Statistical Learning Theory (2nd Edition). Springer, 2000.

[5] Tsoukalas y Uhrig. Fuzzy and Neural Approaches in Engineering. John Wiley and Sons, N.Y., 1997.

[6] Burges, C. A Tutorial on Suppor Vector Machine for Pattern Recognition. Data Mining and Knowledge Discovery. 1998.

[7] Betancourt, Gustavo. Las Máquinas de Soporte Vectorial (SVMs). Scientia Et Technica, vol. XI, núm. 27, abril, 2005, Universidad Tecnológica de Pereira.

[8] Suykens, J.A.K y Vandewalle, J. Least Squares Support Vector Machine Classifiers. Neur.Proc.Lett, 1999.

[9] Glover, F, y G.G Kochenberger. Handbook of Metaheuristics. Kluwer Aca-demic Publishers, 2003.

[10] Eberhart, R y Kennedy, J. A new optimizer using particle swarm theory. Proceedings of the sixth International Symposium on Micro Machine and Human Science, 1995.

[11] Venter, G. and Sobieszczanski-Sobieski, J. Proceedings of the 43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. Denver, 2002.

[12] Jun Sun, Bin Feng, and Wenbo Xu. Particle swarm optimization with particles having quantum behavior. in Proc. Cong. Evolutionary Computation, 2004.

[13] KULeuven-ESAT-SCD. Toolbox v1.7. Disponible vía web en http://www.esat.kuleuven.be/sista/lssvmlab/, revisada por última vez el 01 de diciembre de 2010.

[14] Bishop, C. Neural Networks for Pattern Recognition. New York: Oxford University Press Inc, 1995.

[15] N.Cristianini, J.Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000.

[16] Vapnik, V. Statistical learning theory. Wiley, N.Y., 1998.

[17] Parsopoulos, K.E., y M.N. Vrahatis. A unified particle swarm optimization scheme. Proc. Int. Conf. Computational Methods in Sciences and Engineering (ICCMSE2004). VSP International Science Publishers, 2004.

[18] Engelbrecht, A.P. Fundamentals of Computational Swarm Intelligence. John Wiley & Sons, 2006.

[20] Zweig, M, y G Campbell. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Vol. 39. 4 vols. Clinical Chemistry, 1993.

[21] Smith, Lindsay I. A tutorial on Principal Components Analysis. 2002.

[22] Cuadras, Carles M. Nuevos Métodos de Análisis Multivariante. CIDES, CMC Editions, Barcelona 2010.

[23] Hair, J., Anderson, R., Tathan, R. y Black, W. Análisis Multivariante. Madrid, España: Prentice Hall, 2004.

[24] Kerlinger, F. Y Lee, H. Investigación del comportamiento. México: McGraw-Hill, 2002.