

# Introducción a la Ciencia de Datos

Maestría en Ciencias  
de la Computación

Dr. Irvin Hussein López Nava





04

# Clasificación



## 4.5 Support vector machines

A woman with dark hair, wearing an orange sweater, is shown in profile, looking upwards and to the right. She is holding a tablet in her left hand and has her right hand raised, interacting with a large, glowing blue globe. The background is a complex digital interface with various data visualizations, including line graphs, bar charts, and circular progress indicators. The overall color scheme is dominated by blue and orange.

# ¿Qué son las Máquinas de Soporte Vectorial?



# Definiciones

## ChatGPT

Es un algoritmo de aprendizaje automático que busca encontrar hiperplanos de separación óptimos para clasificar datos en espacios de alta dimensión, destacándose por su capacidad para maximizar el margen entre las clases y su robustez frente al sobreajuste.

<https://chat.openai.com/>

## Wikipedia

Es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte.

[https://en.wikipedia.org/wiki/Support\\_vector\\_machine](https://en.wikipedia.org/wiki/Support_vector_machine)

## Gemini

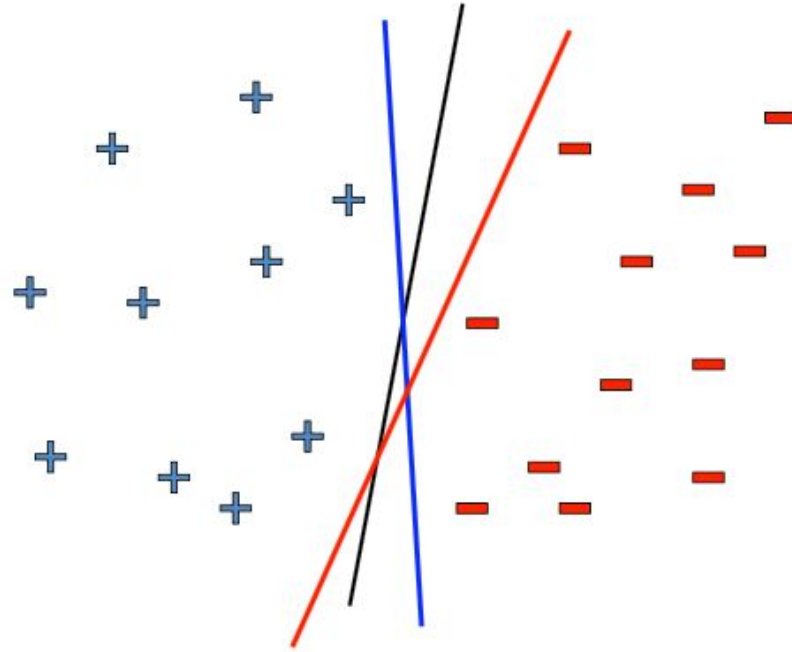
Son un tipo de algoritmo de aprendizaje automático que se utiliza para problemas de clasificación y regresión.

Se basan en el concepto de encontrar un hiperplano que separe los datos de dos clases de forma óptima.

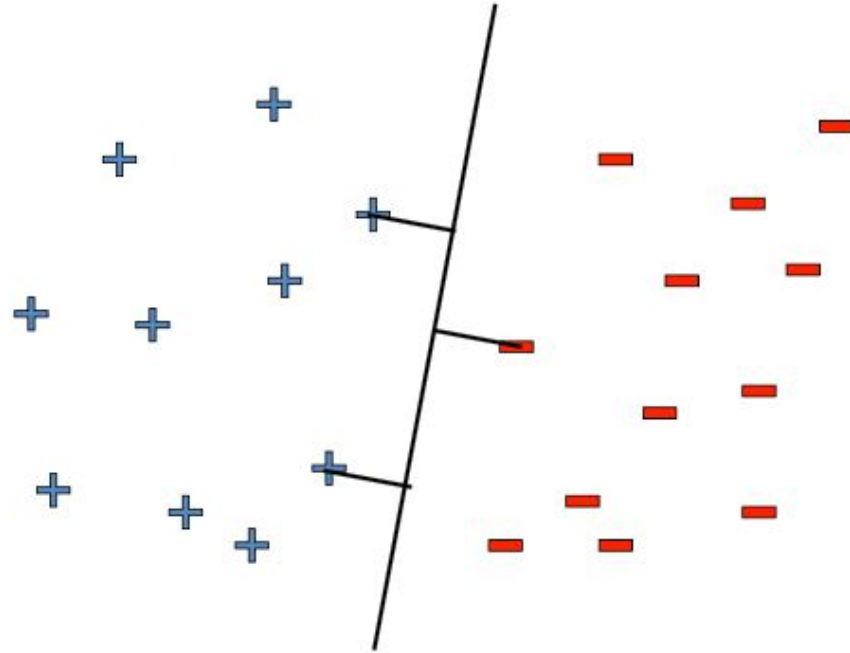
El hiperplano óptimo es el que tiene el margen más amplio entre los dos conjuntos de datos.

<https://gemini.google.com/>

# Clasificadores lineales: ¿qué línea es mejor?



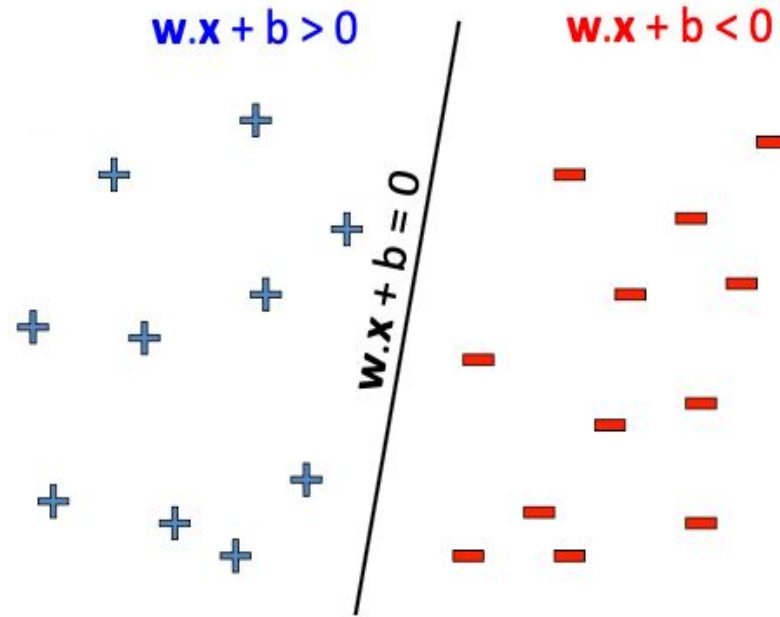
## Elegir el que tenga mayor margen





# Parametrización del límite de decisión

- Confianza:  
 $(w \cdot x_j + b) y_j$
- Etiquetas (class):  
 $y_i \in \{-1, +1\}$



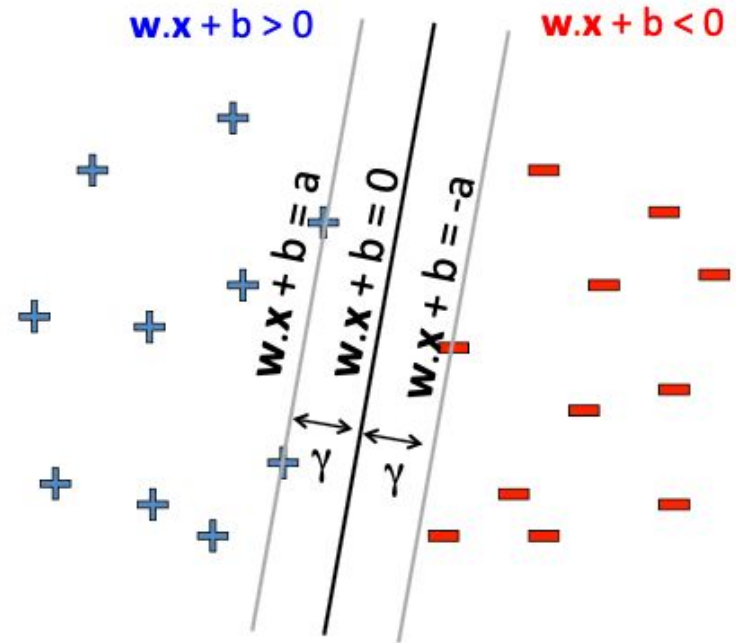
# Maximizando el margen

- Distancia de ejemplos más cercanos desde la línea/hiperplano.  
margen =  $\gamma$

$$\max_{w,b} \gamma = a / \|w\|$$

$$(w \cdot x_j + b)y_j \geq a \quad \forall j$$

- Tener en cuenta que 'a' es arbitraria  
(¿normalizar ecuaciones por a?)



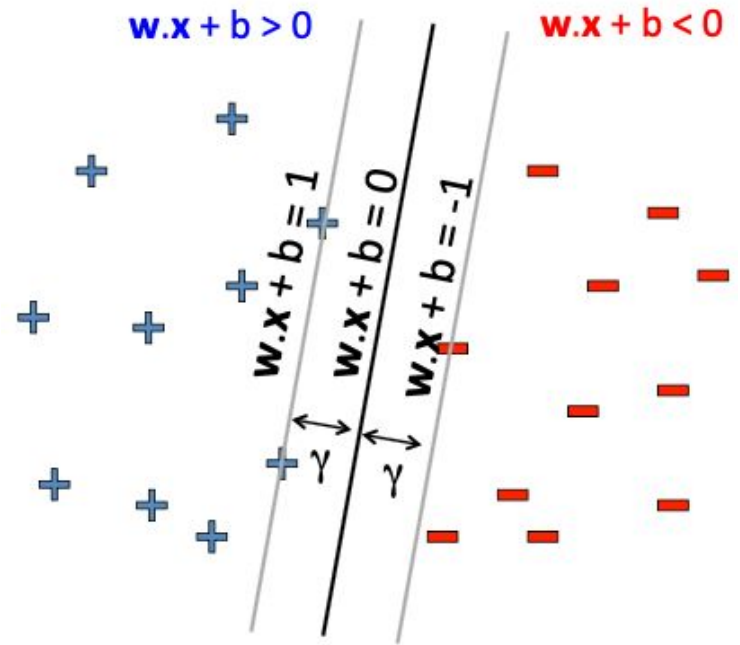
# Máquinas de Soporte Vectorial

- Forma primaria:

$$\min_{w,b} w \cdot w$$

$$(w \cdot x_j + b)y_j \geq a \quad \forall j$$

- Solución eficiente mediante programación cuadrática (QP).
- Algoritmos de solución bien estudiados.



# Formas primarias y duales

- **Forma primaria:** resolver para  $w, b$

$$\min_{w,b} w \cdot w \quad y_l(w \cdot x_l + b) \geq 1 \quad \forall l \in \text{training examples}$$

Prueba de clasificación para nuevos  $x$ :  $w \cdot x + b > 0$

- **Forma dual:** resolver para  $\alpha_1, \dots, \alpha_M$

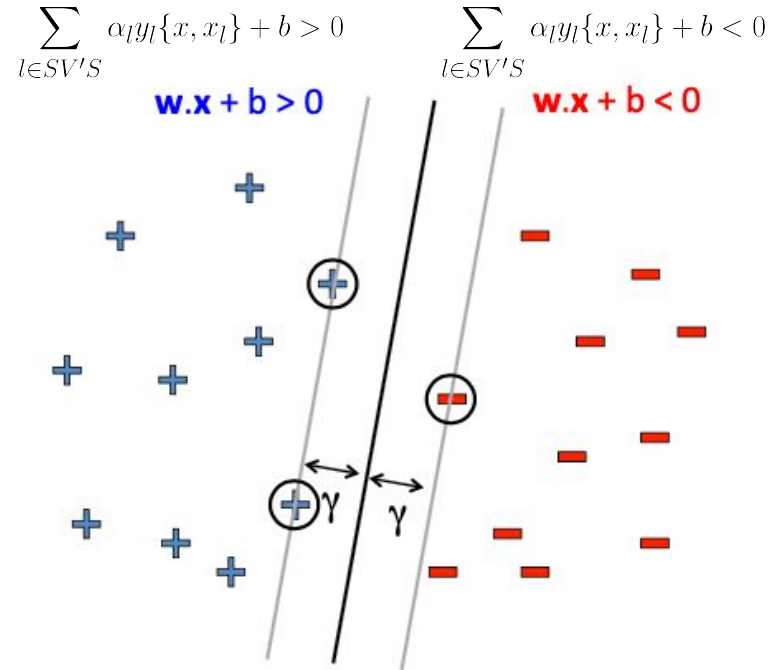
$$\max_{\alpha_1, \dots, \alpha_M} \sum_{j=1}^M \alpha_j - \frac{1}{2} \sum_{j=1}^M \sum_{k=1}^M \alpha_j \alpha_k y_j y_k \{x_j, x_k\} \quad \alpha_l \geq 0 \quad \forall l \in \text{training examples}$$

Prueba de clasificación para nuevos  $x$ :  $\sum_{l \in SV'S} \alpha_l y_l \{x, x_l\} + b > 0$

- ¡Ambos son problemas de QP con un único óptimo local!

# Vectores de soporte

- El hiperplano lineal se define por "vectores de soporte".
- Mover un poco otros puntos no afecta el límite de decisión.
- Sólo es necesario almacenar los vectores de soporte para predecir etiquetas de nuevos puntos.



# Kernel SVM

Y como la forma dual depende sólo de los productos internos, se puede aplicar el truco del núcleo para trabajar en un espacio proyectado (virtual)  $\Phi : X \rightarrow F$ .

**Forma primaria:** resolver para  $w, b$  en el espacio dimensional superior proyectado

$$\min_{w,b} w \cdot w \quad y_l(w \cdot \Phi(x_l) + b) \geq 1 \quad \forall l \in \text{training examples}$$

Prueba de clasificación para nuevos  $x$ :  $w \cdot \Phi(x) + b > 0$

**Forma dual:** resolver para  $\alpha_1, \dots, \alpha_M$  en el espacio de baja dimensión original

$$\max_{\alpha_1, \dots, \alpha_M} \sum_{j=1}^M \alpha_j - \frac{1}{2} \sum_{j=1}^M \sum_{k=1}^M \alpha_j \alpha_k y_j y_k \kappa(x_j, x_k) \quad \alpha_l \geq 0 \quad \forall l \in \text{training examples}$$

Prueba de clasificación para nuevos  $x$ :  $\sum_{l \in \text{SV'S}} \alpha_l y_l \kappa(x, x_l) + b > 0$

# ¿Qué pasa si los datos no son separables linealmente?

Utilizar características de características de características de características...

$$x_1^2, x_2^2, x_1, x_2, \dots, \exp(x_1)$$

- ¡Pero con el riesgo de sobreajuste!

Permitir “error” en la clasificación.

$$\min_{w,b} w \cdot w + C$$
$$(w \cdot x_j + b)y_j \geq 1 \quad \forall j$$

- Maximizar el margen y minimizar el número de errores en los datos de entrenamiento (C – parámetro de compensación).
- No QP
- Pérdida 0/1 (no distingue entre cuasi accidente y error grave).



# Soft-margin SVM

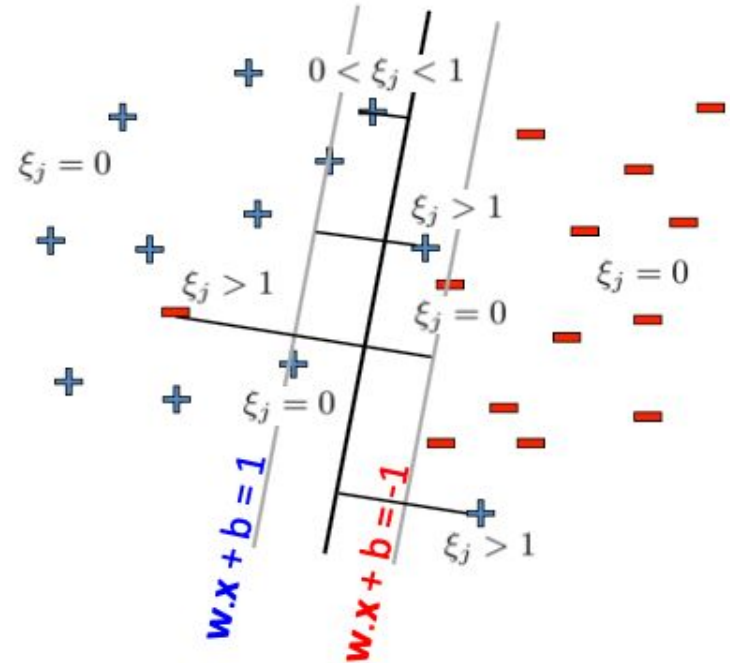
- Permitir "error" en la clasificación.

$$\min_{w,b,\xi} w \cdot w + C \sum \xi_j$$

$$(w \cdot x_j + b)y_j \geq 1 - \xi_j \quad \forall j$$

$\xi_j$  slack variables =  $\xi_j \geq 0 \quad \forall j$   
( $>1$ ) if  $x_j$  is misclassified

- C se elige mediante validación cruzada.  
Todavía QP.

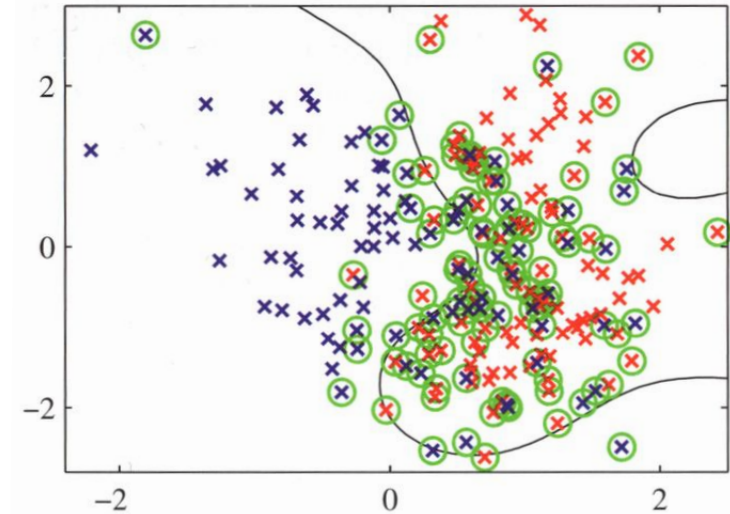


# Soft-margin SVM using Gaussian Kernel

- Los puntos encerrados en un círculo son los vectores de soporte: ejemplos de entrenamiento con  $\alpha$  distinto de cero.
- Puntos trazados en el espacio 2D original.
- Las líneas de contorno muestran constante:

$$\hat{f}(x) = b + \sum_{l=1}^M \alpha_l y_l \kappa(x, x_l)$$

$$\hat{f}(x) = b + \sum_{l=1}^M \alpha_l y_l \exp(-\|x - x_l\|^2 / 2\sigma^2)$$



# Resumiendo SVM

- **Objetivo:** maximizar el margen entre la superficie de decisión y los datos.
- Formulaciones primarias y duales.
  - La dualidad representa la decisión del clasificador en términos de vectores de soporte.
- Kernel SVM.
  - Aprender la superficie de decisión lineal en un espacio de alta dimensión, trabajando en un espacio original de baja dimensión.
- Manejo de datos ruidosos: “variables de holgura (*slack*)” de margen suave
  - De nuevo, formas primarias y duales.
- Algoritmo SVM: optimización de programación cuadrática.
  - Mínimo global único.

# Slack variables - Hinge loss

- Función de pérdida regularizada

$$\xi_j = \text{loss}(f(x_j), y_j)$$

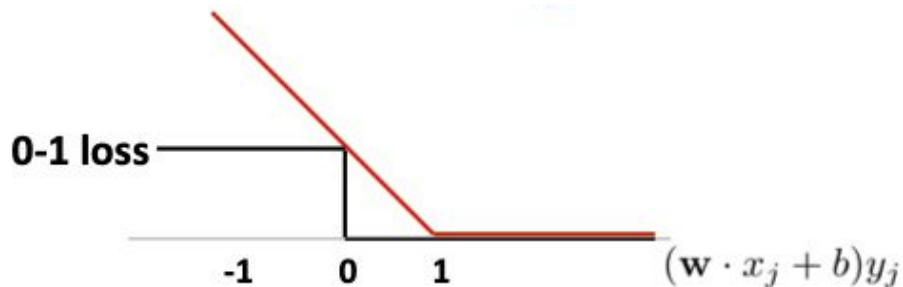
$$f(x_j) = \text{sgn}(w \cdot x_j + b)$$

- *Hinge loss*

$$\xi_j = (1 - (w \cdot x_j + b)y_j)_+$$

Regularization      Loss

$$\min_{w,b,\xi} \underbrace{w \cdot w}_{\text{Regularization}} + C \underbrace{\sum \xi_j}_{\text{Loss}}$$
$$(w \cdot x_j + b)y_j \geq 1 - \xi_j \quad \forall j$$
$$\xi_j \geq 0 \quad \forall j$$



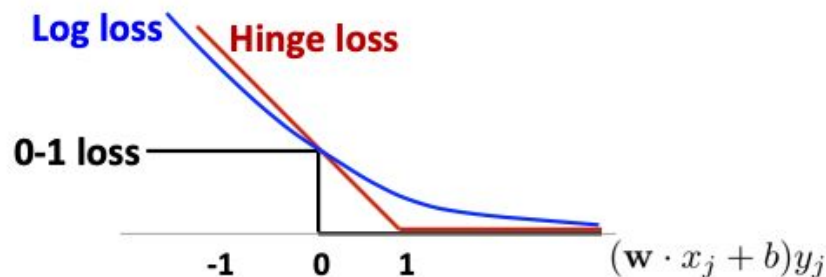
# SVM vs Logistic Regression

- SVM: Hinge loss

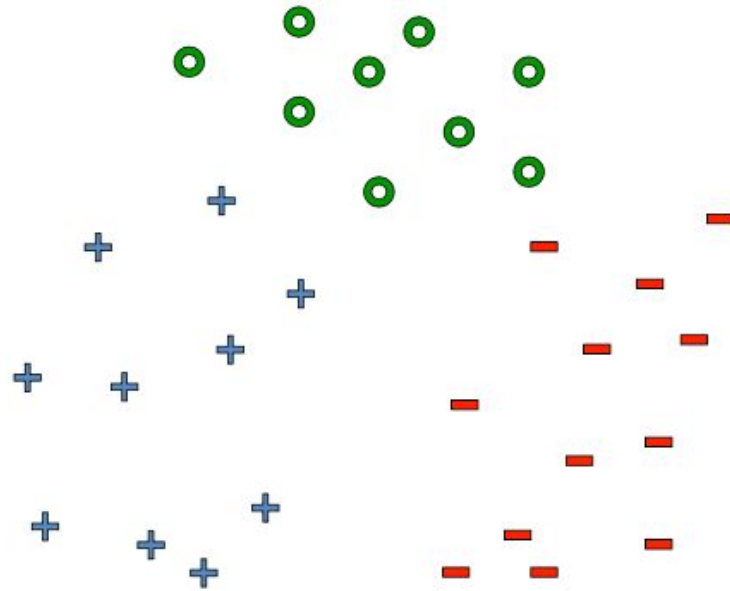
$$\text{loss}(f(x_j), y_j) = (1 - (w \cdot x_j + b)y_j)_+$$

- Logistic Regression: Log loss (conditional likelihood)

$$\begin{aligned}\text{loss}(f(x_j), y_j) &= -\log P(y_j | x_j, w, b) \\ &= \log(1 + e^{-(w \cdot x_j + b)y_j})\end{aligned}$$



## ¿Qué pasa con varias clases?



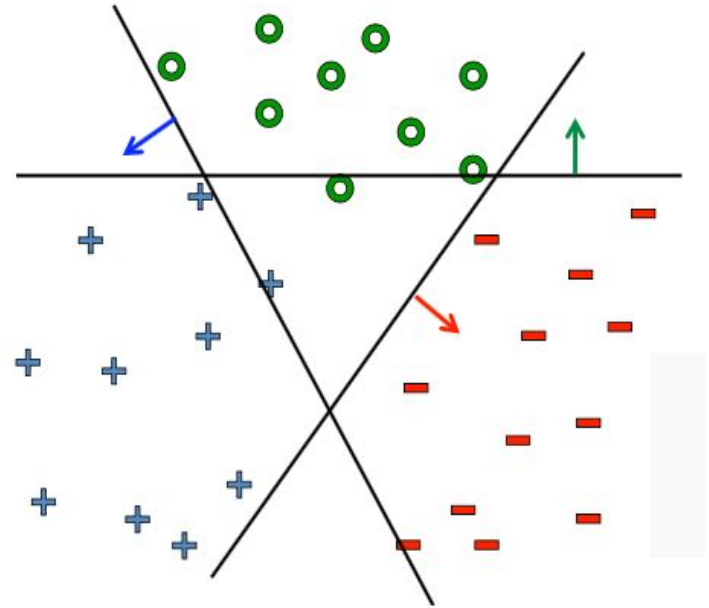
# Solución a problemas multiclase

- Aprender 3 clasificadores por separado:  
Clase k vs el resto

$$(w_k, b_k)_{k=1,2,3}$$

$$y = \arg \max_k w_k \cdot x + b_k$$

- Pero es posible que  $w_k$  no se base en la misma escala.
- Nota:  $(aw).x + (ab)$  también es una solución.





# SVM multiclase

- Aprender simultáneamente 3 conjuntos de pesos:

$$w^{(y_j)}.x_j + b^{(y_j)} \geq w^{y'}.x_j + b^{y'} + 1, \forall y' \neq y_j, \forall j$$

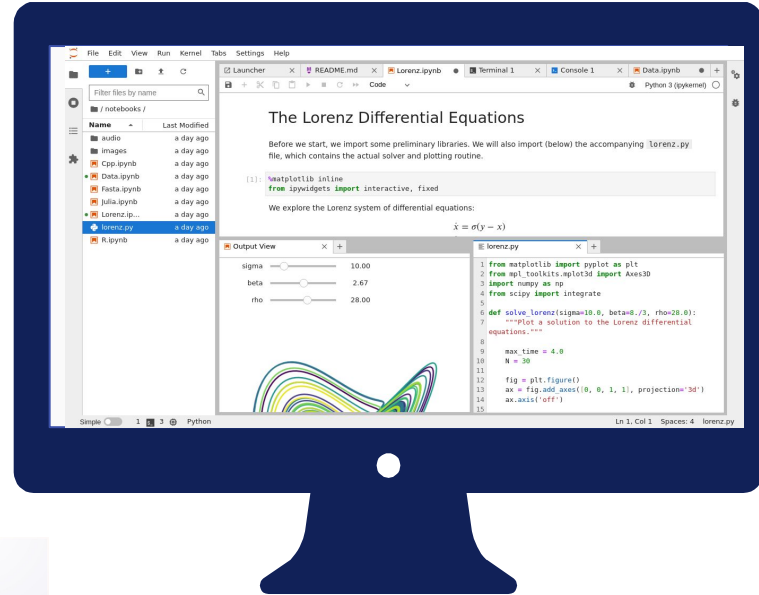
$$\text{minimize}_{w,b} \sum_y w^{(y)}.w^{(y)} + C \sum_j \sum_{y \neq y_j} \xi_j^{(y)}$$

$$w^{(y_j)}.x_j + b^{(y_j)} \geq w^y.x_j + b^y + 1 - \xi_j^{(y)}, \forall y \neq y_j, \forall j$$

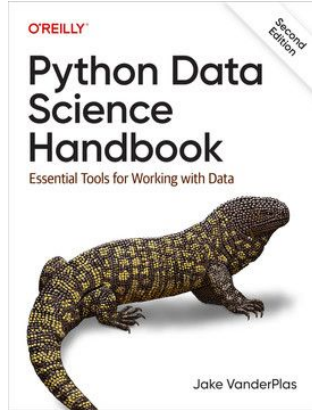
$$\xi_j^{(y)} \geq 0, \forall y \neq y_j, \forall j$$

- $y = \arg \max w^{(k)}.x + b^{(k)}$
- Optimización conjunta:  $w_k$ s tiene la misma escala.

(Go to live notebook)



# Extra Libro



05.07-Support-Vector-Machines.ipynb

# Proyecto

# Evaluación

**20%**

Lecturas y tareas



**50%**

Prácticas



**30%**

Proyecto final

- Las lecturas, tareas y prácticas son obligatorias y deben ser completadas en tiempo y forma.
- Por cada lectura deberá crearse un reporte de máximo una cuartilla.
- Para cada práctica debe crearse un Notebook y deberá agregarse al repositorio personal.
- El reporte del proyecto deberá describir el trabajo de investigación en formato de artículo.

# Requerimientos generales

- El proyecto debe presentar una **propuesta** original, ya sea sobre un tema novedoso o la exploración de un problema poco estudiado.
- Los objetivos del proyecto deben ser alcanzables dentro del tiempo establecido (~un mes) y recursos disponibles.
- La **propuesta de proyecto** debe ser aprobada, en la cual se revisará la pregunta de investigación, datos a utilizar y un esbozo de la metodología propuesta.

# Entregable 1: Reporte de proyecto

## 1. Resumen

- **Objetivo:** Presentar una visión general concisa y clara del proyecto.
- **Contenido:** Problema de investigación, Objetivos específicos; Metodología general; Resultados más relevantes; Conclusiones principales.
- **Consejos:**
  - Debe ser lo suficientemente completo para captar la atención del lector, pero también lo suficientemente conciso para que sea fácil de entender.
  - Se recomienda escribir el resumen al final, una vez que se hayan completado las otras secciones.



# Entregable 1: Reporte de proyecto

## 2. Introducción

- **Objetivo:** Contextualizar el proyecto y presentar el problema de investigación.
- **Contenido:** Relevancia del tema; Justificación de la investigación; Objetivos específicos y pregunta de investigación.
- **Consejos:**
  - La introducción debe ser atractiva y motivar al lector a seguir leyendo.
  - Es importante establecer la relevancia del proyecto para el campo de la ciencia de datos.

# Entregable 1: Reporte de proyecto

## 3. Métodos

- **Objetivo:** Describir en detalle cómo se llevó a cabo la investigación.
- **Contenido:** Conjunto de datos; Procesamiento de los datos; Modelos y algoritmos utilizados; Cómo se evaluarán los modelos.
- **Consejos:**
  - La sección de métodos debe ser lo suficientemente detallada para que otro investigador pueda replicar el estudio.
  - Se deben justificar las decisiones metodológicas tomadas.

# Entregable 1: Reporte de proyecto

## 4. Resultados

- **Objetivo:** Presentar los hallazgos de la investigación de manera clara y concisa.
- **Contenido:** Presentación de los resultados (tablas, gráficos, figuras). Análisis estadístico (si aplica).
- **Consejos:**
  - Los resultados deben ser presentados de manera objetiva (cuantitativa).
  - Se recomienda investigar el tipo de recurso más utilizado en la literatura para presentar los resultados de manera clara y concisa.

# Entregable 1: Reporte de proyecto

## 5. Discusión y conclusiones

- **Objetivo:** Interpretar los resultados y responder a la pregunta de investigación.
- **Contenido:** Interpretación de los resultados; Limitaciones del estudio; Principales hallazgos; Respuestas a las pregunta de investigación; Trabajo a futuro (acciones no alcanzadas).
- **Consejos:**
  - La discusión debe conectar los resultados con la introducción y la problemática abordada.
  - Se deben discutir las limitaciones del estudio y proponer futuras líneas de investigación.

# Entregable 1: Reporte de proyecto

## Referencias

- **Objetivo:** Citar todas las fuentes utilizadas en el proyecto.
- **Contenido:** Lista completa de las referencias bibliográficas; Seguir un estilo de citación específico.
- **Consejos:**
  - Las referencias deben ser formateadas de manera consistente y precisa.
  - Se recomienda utilizar un gestor de bibliografías para facilitar la creación de las referencias (BibTeX).

# Entregable 1: Reporte de proyecto

## Detección de depresión y ansiedad en adolescentes a través de textos de redes sociales

Scarlett Magdaleno Gatica

[scarlett@cicese.edu.mx](mailto:scarlett@cicese.edu.mx)

November 23, 2023

### 1. RESUMEN

Este estudio se centra en explorar la comorbilidad entre la ansiedad y la depresión en adolescentes a través del análisis de datos recopilados de comentarios en redes sociales. Para ello, se ha desarrollado un clasificador de texto utilizando técnicas de procesamiento del lenguaje natural y aprendizaje automático. Los resultados obtenidos revelan descubrimientos significativos en relación con la conexión entre problemas de sueño, ansiedad y depresión. Además, se lleva a cabo un análisis exhaustivo para determinar la representación óptima de los datos, ya sea mediante unigramas, bigramas o trigramas. Los resultados del clasificador sugieren que este enfoque muestra un prometedor potencial para la detección temprana de trastornos mentales en la población adolescente.

### 2. INTRODUCCIÓN

La ansiedad y la depresión son trastornos frecuentes entre los jóvenes, con consecuencias significativas para la salud mental y el bienestar. Estas enfermedades son altamente comórbidas, es decir, se presentan juntas, y la ansiedad patológica es una precursora regular en el desarrollo de la depresión [1].

La depresión, en particular, se asocia con consecuencias graves y se ha identificado como la principal causa de suicidio. En adolescentes de 12 a 19 años, el suicidio ocupa el tercer lugar como causa de muerte [2]. Este sombrío panorama subraya la importancia de abordar estos trastornos en una etapa temprana.

Las redes sociales se han convertido en una parte integral de la vida diaria de los jóvenes, proporcionando una oportunidad única para identificar posibles indicadores de ansiedad y depresión mediante el análisis de los textos que comparten. Además de los cambios en el estado de ánimo, existen síntomas más evidentes como alteraciones en el apetito, pérdida de energía e insomnio [3]. Este estudio busca aprovechar estas señales indirectas presentes en los textos de redes sociales para detectar indicios de ansiedad y depresión en comentarios realizados por jóvenes y adolescentes.

En este contexto, surge la necesidad de desarrollar herramientas que permitan la detección temprana de depresión y ansiedad en adolescentes para brindar tratamiento y prevenir consecuencias graves como el

suicidio. Este estudio tiene como objetivo desarrollar un clasificador de texto capaz de detectar si un comentario de redes sociales hecho por un adolescente de 13 a 17 años indica la presencia de ansiedad y depresión.

En este marco propuesto, se emplearán técnicas de procesamiento del lenguaje natural (PLN) y seis algoritmos de aprendizaje automático para detectar la comorbilidad de ansiedad y depresión. El modelo utiliza PLN para identificar patrones e indicadores de ansiedad y depresión en textos escritos, como comentarios o mensajes en redes sociales.

### 3. MÉTODOS

#### 3.1. Conjunto de Datos

El conjunto de datos titulado *Student-Depression-Text* utilizado en este estudio se obtuvo de Kaggle y fue generosamente proporcionado por Nidhi-Yadav [4]. Este dataset consta de 7489 instancias y cinco atributos clave que detallan información valiosa para nuestro análisis:

- **Text:** Contiene datos en inglés provenientes de comentarios de diversas redes sociales.
- **Labels:** Representa la etiqueta de clasificación, siendo "0" indicativo de la ausencia de ansiedad y depresión, mientras que "1" señala la presencia de estos trastornos.
- **Age:** Incluye edades que abarcan el rango de 13 a 17 años.
- **Age Category:** Divide las edades en dos categorías: "Young Age" (13-15 años) y "Teen Age" (16-17 años).
- **Gender:** Especifica el género del individuo, clasificado como Masculino o Femenino.

La metodología utilizada se describe a continuación.

#### 3.2. Análisis Exploratorio, Limpieza y Preprocesamiento de Datos

Se llevó a cabo un análisis exploratorio exhaustivo para comprender la naturaleza y las características del conjunto de datos. Este análisis reveló desequilibrios de clases, edades etiquetadas incorrectamente y tendencias en la prevalencia de ansiedad y depresión según el género y la edad.

La limpieza de datos fue una etapa esencial para garantizar la calidad y eficacia de los modelos. Se aplicaron diversas técnicas, incluida la reducción de datos

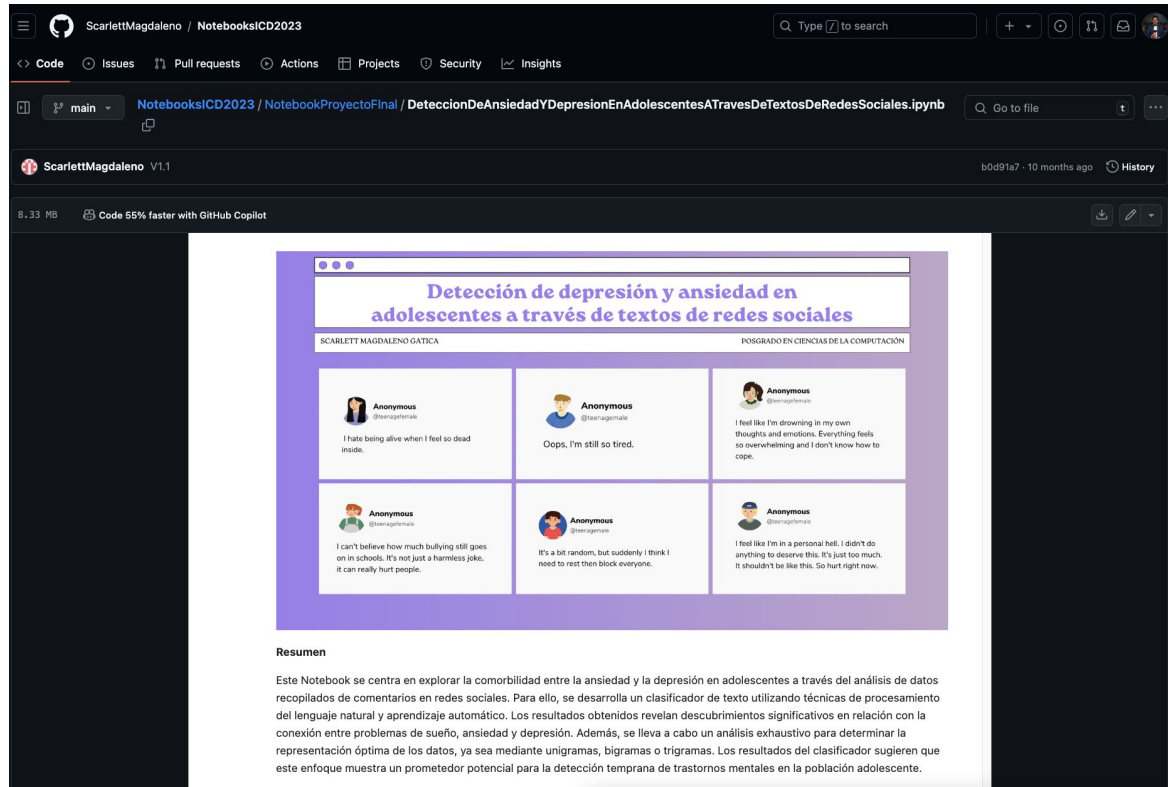
## Entregable 2: Notebook del proyecto

- El proyecto deberá ser desarrollado usando un Notebook de Python, el cual deberá estar organizado y documentado.
- Verificar que las salidas (impresiones) sean las adecuadas y evitar mostrar salidas útiles solo para el periodo de desarrollo.
- Se deberá proporcionar el enlace al repositorio de Github personal.

Cronológicamente es el primer elemento de trabajo del Proyecto (solo después de la propuesta).



# Entregable 2: Notebook del proyecto



ScarlettMagdaleno / NotebooksCD2023

Code Issues Pull requests Actions Projects Security Insights

NotebooksCD2023 / NotebookProyectoFinal / DeteccionDeAnsiedadYDepresionEnAdolescentesATravesDeTextosDeRedesSociales.ipynb







ScarlettMagdaleno V1.1

b0d91a7 - 10 months ago History

8.33 MB Code 55% faster with GitHub Copilot

### Detección de depresión y ansiedad en adolescentes a través de textos de redes sociales

SCARLETT MAGDALENO GATICA POSGRADO EN CIENCIAS DE LA COMPUTACIÓN

|  |  |  |
|--|--|--|
| <br><b>Anonymous</b><br>@scarlettma1<br>I hate being alive when I feel so dead inside.  | <br><b>Anonymous</b><br>@scarlettma1<br>Oops, I'm still so tired.   | <br><b>Anonymous</b><br>@scarlettma1<br>I feel like I'm drowning in my own thoughts and emotions. Everything feels so overwhelming and I don't know how to cope.                    |
| <br><b>Anonymous</b><br>@scarlettma1<br>I can't believe how much bullying still goes on in schools. It's not just a harmless joke, it can really hurt people. | <br><b>Anonymous</b><br>@scarlettma1<br>It's a bit random, but suddenly I think I need to rest then block everyone. | <br><b>Anonymous</b><br>@scarlettma1<br>I feel like I'm in a personal hell. I didn't do anything to deserve this. It's just too much. It shouldn't be like this. So hurt right now. |

#### Resumen

Este Notebook se centra en explorar la comorbilidad entre la ansiedad y la depresión en adolescentes a través del análisis de datos recopilados de comentarios en redes sociales. Para ello, se desarrolla un clasificador de texto utilizando técnicas de procesamiento del lenguaje natural y aprendizaje automático. Los resultados obtenidos revelan descubrimientos significativos en relación con la conexión entre problemas de sueño, ansiedad y depresión. Además, se lleva a cabo un análisis exhaustivo para determinar la representación óptima de los datos, ya sea mediante unigramas, bigramas o trigramas. Los resultados del clasificador sugieren que este enfoque muestra un prometedor potencial para la detección temprana de trastornos mentales en la población adolescente.

## Entregable 3: Presentación del proyecto

- A diferencia de los entregables previos, la presentación permitirá dar a conocer la investigación realizada por cada estudiante al resto de la clase, por lo que se deben ser claros, concisos y amenos.
- El número de *Slides* para la presentación deberá ser 12 incluyendo la portada, referencias, y agradecimientos. Después de la portada deberá mostrarse “el video”.
- La fecha para las presentaciones es el miércoles 20/nov, y cada estudiante tendrá un máximo de 10 minutos.
- La secuencia de las presentaciones seguirá un orden inverso al de recepción de los trabajos.

# Entregable 3: Presentación del proyecto



**Detección de depresión y ansiedad en adolescentes a través de textos de redes sociales**

Posgrado en Ciencias de la Computación  
Estudiante: Scarlett Magdaleno Gatica

## Entregable 4: Video del proyecto

- Preparar un video de 1 minuto (exacto) detallando lo realizado en el proyecto.
- Para este entregable se evaluará el nivel de abstracción y síntesis del proyecto, así como la creatividad y calidad.
- El video debe ser dirigido a un objetivo no técnico: formato de divulgación de la ciencia.
- El video también deberá motivar a nuevos estudiantes a tomar el curso :)

Cronológicamente es el último elemento a preparar.

## Entregable 4: Video del proyecto



# Gracias!

¿Alguna pregunta?

hussein@cicese.mx

<https://sites.google.com/view/husseinlopeznava>



**CREDITS:** This presentation was based on a template by [Slidesgo](#), and includes icons by [Flaticon](#).