UNIVERSIDAD DE PIURA

FACULTAD DE INGENIERÍA

Estimación de la plaga de trips de la mancha roja en el banano orgánico mediante técnicas de *Machine Learning*

Tesis para optar el Título de
Ingeniera Mecánico - Eléctrica

Frescia Soledad Sernaqué Ramos

Asesor:
Mgtr. Ing. José José Manrique Silupú

Piura, julio de 2022

UNIVERSITAS STUDIORUM PIURENSIS

Dedicatoria

A mis padres Lazaro, Cecilia y a mi hermana Ingrid, porque a pesar de todas las dificultades siempre son mi apoyo incondicional, y a Cristhian por su paciencia, amor y apoyo en este arduo proceso.

UNIVERSITAS STUDIORUM PIURENSIS

Agradecimientos

El presente trabajo de investigación fue ejecutado en el marco del subproyecto "Transformación Digital del sector agro – industrial aplicado al banano orgánico", financiado por Banco Mundial y el Consejo Nacional de Ciencia, Tecnología e Innovación – CONCYTEC, bajo el contrato N°165 – 2018 – FONDECYCT – BM – IADTAV, en colaboración con la Universidad de Piura (UDEP) y la cooperativa ASPROBO.

UNIVERSITAS STUDIORUM PIURENSIS

Resumen

La presente investigación está enfocada en solucionar uno de los mayores problemas de la producción del banano orgánico, uno de las principales frutas peruanas de exportación, bien valorada en Europa, América del Norte y Centroamérica. Esta investigación presenta la relación entre los factores climáticos y el crecimiento y/o reproducción del trip de la mancha roja, una plaga que amenaza al banano orgánico constantemente.

Para llevar a cabo esta investigación, se estudiaron tres técnicas de aprendizaje no supervisado de Machine Learning: regresión lineal, *support vector machine* y *Twin Support Vector Machine (TSVM)*. Al tratarse de valores reales, se determinó que la TSVM fue la técnica que más se aproximó al comportamiento del trip en base a los cambios de los factores climáticos del lugar.

Los factores climáticos se eligieron, en primer lugar, en base al criterio de investigaciones realizadas en otros países exportadores de productos como café, frutas, entre otros; posteriormente, con la información proporcionada directamente por los agricultores de banano orgánico, se realizaron las evaluaciones y correcciones de estos factores.

Cabe indicar que, la tesis ha sido realizada, de forma experimental, en una de las parcelas de la Cooperativa ASPROBO, ubicada en el distrito de Buenos Aires de Morropón, provincia de Piura, y los avances y resultados obtenidos han ayudado al propietario a mejorar sus técnicas y criterios de evaluación de sus cultivos.

UNIVERSITAS STUDIORUM PIURENSIS

Tabla de contenido

Lista de figuras

Lista de tablas

UNIVERSITAS STUDIORUM PIURENSIS

Introducción

La presente investigación ha sido elaborada con la finalidad de hallar una correlación entre una parcela de banano orgánico y el desarrollo del trip de la mancha roja, una de las plagas que afecta a estas plantaciones. Para lograr esta correlación, se ha recurrido a las técnicas del Internet de las Cosas, que, a través de las técnicas de aprendizaje supervisado, permiten integrar las variables recolectadas.

Como punto de partida para esta investigación, se seleccionó una parcela de la Cooperativa ASPROBO (Asociación de pequeños productores de banano orgánico), que se ubica en el distrito de Buenos Aires de Morropón, en la región Piura. De esta parcela, se recolectó información real sobre la presencia y control de plagas, que representan un problema para la exportación del banano orgánico. Además, se instaló una estación meteorológica y unos sensores o nodos para recolectar datas agroclimáticas.

En el primer capítulo de la investigación se describe el proceso de cultivo del banano orgánico y la importancia que representa para los pequeños agricultores de Buenos Aires de Morropón, quienes se han organizado en cooperativas para facilitar la exportación de este producto. Además, se describe la problemática en torno a la presencia del trip de la mancha roja.

En el segundo capítulo, se revisan las técnicas de aprendizaje supervisado que permiten procesar la data obtenida durante la investigación, a través de un algoritmo, para predecir la aparición del trip de la mancha roja en las plantaciones de banano. De las técnicas analizadas, se precisa cuál fue resulta más apropiada, de acuerdo a las características de la parcela piloto en ASPROBO.

En el tercer capítulo, se detalla cómo se obtuvo la información de la plaga y las variables agroclimáticas seleccionadas. Del mismo modo, se describen las características y funcionamiento de la estación meteorológica y unos sensores utilizados.

Finalmente, en el cuarto capítulo, se presentan los resultados obtenidos de la simulación y aplicación del Twin Support Vector Machine (TSVM) para predecir la aparición del trip de la mancha roja en las plantaciones de banano orgánico en ASPROBO.

UNIVERSITAS STUDIORUM PIURENSIS

Capítulo 1
Marco teórico

Existen dos tipos de banano que son cultivados en Perú: el banano convencional y el banano orgánico. A diferencia del banano convencional, el proceso de siembra y cosecha del banano orgánico requiere más cuidado e inversión para alimentar los suelos, controlar las plagas, cosechar el producto, entre otros factores. Sin embargo, las ganancias que se obtienen son mayores a las ganancias obtenidas de la importación del banano convencional, por ello, algunos productores optan por dedicarse al cultivo de este producto; esto es debido al incremento de exportación del banano orgánico en los últimos años, es decir, se pasó de una exportación equivalente a US$119 millones en el 2014 a una exportación equivalente a US$152 millones en el 2016 (Instituto de Investigación y Desarrollo de Comercio Exterior de la Cámara de Comercio de Lima - IDEXCAM, 2017). No obstante, su producción no se encuentra exenta de plagas y peligros.

1.1 Banano orgánico

Una de las mayores preocupaciones por mantener un estilo de vida saludable y alimentar correctamente el cuerpo, ha permitido que los productos orgánicos sean muy valorados en la sociedad. El banano orgánico es una de esas valiosas alternativas alimentarias, pero, además, en el Perú, representa una importante fuente de ingresos económicos para las familias dedicadas a su agricultura.

Las zonas de producción más grandes de banano orgánico en el Perú se encuentran en la costa norte, en las regiones de Tumbes, Piura y Lambayeque, donde el clima cálido y las lluvias moderadas resultan indispensables para su crecimiento (Agricultura, 2017). En estas zonas, los agricultores poseen entre 1 y 3 hectáreas y se encuentran organizados en asociaciones o cooperativas.

En el 2008, se contabilizó un aproximado de 3 414 hectáreas de banano orgánico certificadas, de las cuales, el 80% están en Piura (Reinoso, 2008). En tanto, para el 2019, ya existían 10 500 hectáreas (Rosales, 2019), lo cual representa una tendencia de crecimiento positivo.

El mercado de este producto se encuentra principalmente en territorio extranjero, tal es así que, en los últimos años, Perú ha logrado unirse a la fila de los países exportadores de

banano orgánico a Europa y Norteamérica (Minagri: Exportación de banano orgánico peruano creció 94% en últimos 5 años, 2015).

1.2 Ciclo del banano

El banano es una planta que no requiere un proceso de siembra continuo, es decir, no se tiene que sembrar cada año para obtener frutos. Asimismo, el banano es una planta que suele generar sus propios hijuelos, los cuales, con el tiempo, tendrán su propio fruto y sus propios hijuelos.

No obstante, esta constante cadena puede tener variaciones de patrón o ciclo, según la forma que eligen sus agricultores (Agricultura, 2017). A continuación, se presenta el patrón que caracteriza a los agricultores piuranos.

1.2.1 *Selección y preparación de terreno*

Lo ideal para los agricultores es elegir terrenos cercanos a alguna vía, canal o acceso de agua; sin embargo, estos canales no están al alcance de todos, por lo que, el agua se convierte en uno de los principales problemas y condiciones a tomar en cuenta para su producción. Para intentar revertir esta situación, se sugiere tomar medidas topográficas que permitan localizar o crear un sistema de drenaje para la plantación, además de los análisis de suelo que son importantes para el óptimo crecimiento del banano (Ing. Juan Carlos Rojas Llanque, 2012).

1.2.2 *Hoyado y selección de semilla*

A continuación, se realiza el proceso de hoyado y selección de semilla. El hoyado que se recomienda realizar es de 0.4mx0.4mx0.4m; esto con la finalidad de no afectar la producción de la planta durante los periodos de sequía que se puedan presentar. A modo de recomendación, se les aconseja realizar el hoyado, días antes de la siembra, y de esta manera poder abonar bien el suelo, esta mezcla de abono suele colocarse al fondo del hoyo, y así se asegura que las raíces estarán bien alimentadas. (Ing. Juan Carlos Rojas Llanque, 2012)

Para la semilla, se eligen aquellas que provienen de lugares certificados orgánicamente. Por otro lado, existe la opción de colocar plantaciones con semanas de crecimiento, el cual favorece el proceso de producción del fruto.

1.2.3 *Siembra*

Las semillas y/o plantaciones son colocadas en forma vertical en el lugar previamente seleccionado y hoyado, el cual aportará orden en el crecimiento de la planta y el fruto.

1.2.4 *Deshermane y eliminación de brotes*

El banano es una planta que suele producir entre 5 a 8 hijuelos (ver figura 1), y es ahí donde el productor, según su criterio, mantiene el que tiene mejor posición o crecimiento y los otros deben ser eliminados.

El proceso de deshermane se debe realizar a las 10 semanas o 2.5 meses y la eliminación de rebrotes, a las 12 semanas o 3 meses de crecimiento de la planta madre.

Los hijuelos pueden ser elegidos entre las semanas 10 y 14 y en la semana 20, donde el agricultor confirma si su criterio de elección fue el correcto. De lo contrario, puede eliminarlo y dejar a otro que tenga mejor posición.

Figura 1. Hijuelos alrededor de la planta del banano

Por el contrario, tal como se puede apreciar en la figura 1, si los hijuelos no necesarios no son removidos, estos consumen los minerales que necesita el hijuelo seleccionado, afectando su crecimiento y el de la planta madre.

1.2.5 *Riego*

Según los estudios realizados, el tiempo ideal para regar las plantaciones es cada 15 días en verano o 20 días en invierno; sin embargo, en el norte del país los periodos de sequía suelen ser largos; por lo tanto, los agricultores siguen otro proceso, que está en función a la disponibilidad de agua (según orden de llegada) que almacenan en pozos creados por la asociación o cooperativa.

1.2.6 *Control de malezas*

Las malezas son todas las hojas secas que caen de las plantaciones y que con el tiempo se van acumulando en el suelo (ver figura 2), favoreciendo la reproducción acelerada de plagas. Esta debe ser retirada periódicamente, pero sin aplicación de productos químicos.

Figura 2. Malezas en plantaciones de banano

1.2.7 *Control de plagas y fertilización*

Las plagas varían según la localidad o región. Por ejemplo, en el distrito de Buenos Aires, Morropón, existen nueve tipos de plagas comunes y recurrentes, siendo el trip de la mancha roja la más peligrosa. Esta plaga produce unas manchas negras en la cáscara del fruto, lo que trae como consecuencia, que el banano sea descartado por los compradores extranjeros.

Para combatir esta plaga y, al mismo tiempo, reforzar el crecimiento del banano, los productores pueden elegir entre cinco fertilizantes orgánicos.

Tabla 1. Tipos de abono orgánicos

Tipo de abono	Ha / año		N	P	K	Ca	Mg	S
	Sacos	Kg	Kg/ha	Kg/ha	Kg/ha	Kg/ha	Kg/ha	Kg/ha
Compost	180	6300	75	24	15	0	0	0
Sulfato de potasio	15	750	0	0	375	0	0	135
Guano de Islas	58	2900	348	203	58	232	14	43
Kimelgran	4	100	1	0	4	1	2	0
Sulphomag	9	450	0	0	99	0	81	99
Total	266	1500	424	227	551	233	97	277

Los fertilizantes se colocan a criterio del productor, quien realiza un estudio preliminar de la zona, como la falta de agua, la existencia de plagas, las condiciones del suelo, entre otros factores.

Para el caso del control de plagas, los productores de ASPROBO suelen utilizar productos naturales: sal, ají (como fungicida orgánico) y otros productos orgánicos aprobados por la FAO[1]. Cabe mencionar que, los productores de Tumbes, Piura y Lambayeque recurren a diferentes criterios y alternativas de control de plagas.

1.2.8 *Pre – cosecha del banano*

Antes de la cosecha, el banano pasa por distintos procesos, los cuales se detallan a continuación:

- Embolse prematuro. Se realiza para censurar la presencia de insectos que pretendan alimentarse del fruto. Este embolsa se ubica en la bacota recién brotada de la planta (ver figura 3).

Figura 3. Bacota del banano

Fuente: Cosecha, corte y empaque del banano convencional y orgánico para exportación (Chumbes).

- Desflore. Este proceso requiere de observar el crecimiento del banano y cuando las manos de banano se colocan en posición hacia arriba, se deben retirar las flores, que por lo general son blandas.

[1] FAO es la Organización de las naciones para la agricultura y la alimentación

Figura 4. Derecha - presencia de flor en la mano de banano / Izquierda – embolse
Fuente: Cosecha, corte y empaque del banano convencional y orgánico para exportación (Chumbes).

- Desmane. El racimo del banano suele tener muchas manos, las cuales, no necesariamente son importantes o vitales en el racimo, ya que se ha demostrado, que las manos más pequeñas o que no están en orden, deben ser retiradas para que las demás manos tengan espacio suficiente para desarrollarse y crecer hasta el tamaño necesario para ser exportadas. Sin embargo, al retirar estas manos, se suele dejar un dedo de la última mano, con la finalidad de engañar a los insectos (en caso alguno llegue al embolse, se encontrará con el dedo trampa y se quedará ahí mismo, sin necesidad de seguir subiendo y malograr todo el racimo).

- Corte de dedos laterales. Así como en el desmane, el corte de dedos laterales es el corte de dedos extremos de las manos, el cual tiene la función de permitir que no se desperdicie en la caja de empaque. Los compradores tienen ciertas exigencias, una de ellas es la medida mínima de largo y diámetro de los bananos; es por ello, que se eliminan los dedos extremos, ya que al dejarlos quitarían espacio y no crecerían lo suficiente para que sean empacados para exportación.

- Desbacote. En este proceso se retiran las flores “masculinas”, que se envuelven en la bráctea, permitiendo que esta se alargue un poco (ver figura 5).

Figura 5. Proceso de desbacote
Fuente: Cosecha, corte y empaque del banano convencional y orgánico para exportación (Chumbes).

- Uso de cintas. Luego de realizar todos los puntos antes mencionados, se procede a identificar el racimo con un color de cinta, con la finalidad que en once o doce semanas sea cortado (ver figura 6).

Figura 6. Uso de cintas de colores
Fuente: Cosecha, corte y empaque del banano convencional y orgánico para exportación (Chumbes).

- Uso de cuello de monja. Los cuellos de monja se colocan entre las manos del racimo del banano, con la finalidad de proteger la punta de los bananos que están en la parte inferior. Este procedimiento permite reducir daños de merma en el proceso de empaque.

Figura 7. *Cuello de monja*

- Amarre. Para evitar que la planta caiga cuando ha dado fruto, por el peso del racimo, es necesario sujetar la planta con cables o sujetadores desde la parte alta de la misma y en posición contraria a donde está el peso, tal como se muestra en la figura 8.

Figura 8. Amarre de la planta para evitar caer
Fuente: Cosecha, corte y empaque del banano convencional y orgánico para exportación (Chumbes).

1.2.9 *Corte y empaque*

Según la tecnología, asociación, departamento o distrito, el corte de racimo es distinto, ya que esto depende del tipo de transporte de racimos hacia la empacadora, la cual puede ser fija o móvil, según sean los recursos de los que se disponen en la plantación. Los casos más conocidos de transporte de racimos son:

- Sistema tradicional: una persona corta a una altura determinada para que la planta se incline y, de esta manera, el racimo pueda bajar despacio y con cuidado hasta la espalda otra persona, la cual lo transportará con cuidado hasta la empacadora.

- Sistema nuevo: una persona corta a una altura determinada para que la planta se incline y, de esta manera, el racimo pueda bajar despacio hasta colocarlo en el riel más cercano, la cual llevará al racimo a la empacadora mediante el cable vía.

Cada sistema presenta las siguientes ventajas y desventajas:

Tabla 2. Cuadro comparativo de sistemas de transporte

	Ventaja	Desventaja
Sistema tradicional	No es costoso	Al ser transportado en la espalda de un ser humano, tiende a sufrir golpes que pueden perjudicar al racimo
Sistema nuevo	El racimo se transporta hasta la empacadora sin daños ni estropeos.	Sistema tecnológico costoso

Independiente del sistema de traslado, en Perú, el racimo se traslada siempre completo, nunca por partes. Luego, al llegar a la empacadora, se le realizan algunos cortes, según el mercado de destino; por ejemplo, en algunos países exigen manos de 5 dedos, mientras que, en otros, exigen de 4 dedos. A este proceso también se le conoce como desmane, pero en la post cosecha del banano.

Para este proceso, se requiere que el cortador sea muy hábil: haciendo uso de cuchillos curvos o cortadores semicirculares debe realizar un corte limpio y sin dejar otros cortes ni desgarres; además el corte debe realizarse muy cerca al tallo dejando suficiente corona.

Figura 9. Corte del racimo
Fuente: Cosecha, corte y empaque del banano convencional y orgánico para exportación (Chumbes).

1.2.10 *Lavado y saneo*

Una vez cortado, el racimo es depositado en tinas para su lavado y saneo. Estas tinas suelen clasificarse según su posición fija o móvil en dos tipos:

- Empacador móvil. Método antiguo que consistía en movilizar una tina grande para armar el sistema de empacadora en distintos lugares.
- Empacadora fija: Método actual, consiste en construir al medio, al costado o en algún extremo, una empacadora que consta de piscinas grandes con sistema de drenaje para darle un lavado correcto a las manos de banano que han sido cortados.

En cualquiera que sea el caso, las empacadoras tienen como finalidad lavar las manos de banano y eliminar las deformaciones o defectos que presenten como estropeo o daño (suelen utilizar cuchillos afilados para poder eliminar dichos estropeos) , ver figura 10.

Figura 10. Lavado de manos del racimo del banano
Fuente: Cosecha, corte y empaque del banano convencional y orgánico para exportación (Chumbes).

1.2.11 *Enjuague o desleche*

Luego de retirar los dedos o las manos del racimo, el banano suele desprender un líquido conocido como “leche”, el cual es removido durante el proceso de lavado, también denominado “desleche”. Para lograr esto, el banano debe ser remojado en agua entre doce y quince minutos.

En la tina de lavado, que presenta divisiones, el banano es separado y clasificado según el número de dedos, de acuerdo al mercado de destino (Ver figura 11).

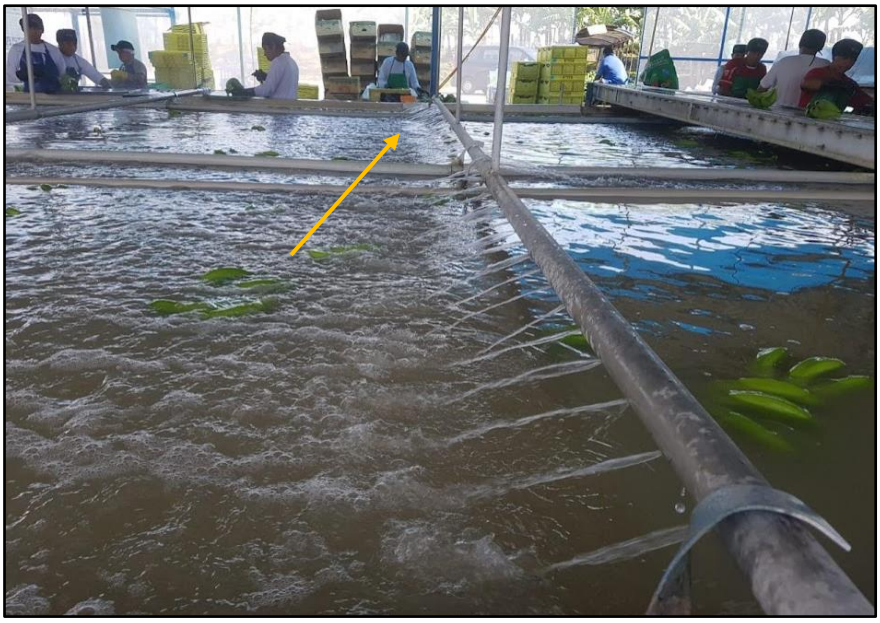

Figura 11. Proceso de enjuague y clasificación

1.2.12 *Sello y pesado*

Una vez limpio, el banano es pesado y clasificado para su comercialización al mercado extranjero. En ese momento, a las manos también se les coloca unas pegatinas que acreditan que se trata de banano orgánico y que, además, contiene información sobre el lugar de origen y detalles de la empresa exportadora.

Cabe indicar que, a las manos de banano que, alcanzado el peso requerido por los compradores, se les añade medio kilo más, con la finalidad de suplir la pérdida de peso que sufre el producto al deshidratarse durante el periodo de transporte, de forma que, a su llegada, la venta de las cajas no se vea perjudicada (Ver figura 12).

Figura 12. Sellado de las manos de banano

1.2.13 *Empaque*

Posteriormente, el banano sellado y pesado se coloca en cajas de cartón corrugado (ver figura 13), según las especificaciones y dimensiones solicitadas por los compradores. Las cajas también poseen información sobre el producto que contienen, tales como el peso, lugar de procedencia, información de la empacadora, fecha, entre otras. Por lo general, las cajas poseen unos agujeros que permiten ventilar la fruta durante su transporte hasta el mercado de destino.

Aunque en Perú se ha incrementado el número de cooperativas y asociaciones de productores de banano, con la finalidad de mejorar la exportación de sus productos, es necesario aclarar que no a todas las asociaciones y cooperativas se les pide los mismos requisitos de exportación y por tanto, no todas cuentan con los equipos o materiales que ayuden a realizar un control eficiente de las plagas (Juan Carlos Rojas Llanque, 2013).

Figura 13. Empaque

1.2.14 *Transporte*

Finalmente, las cajas con banano orgánico son trasladas hasta el destino del comprador en contenedores cerrados, los cuales tienen la función de proteger al producto del polvo, calor (poseen refrigeración) y agua.

1.3 Plagas del banano

El ingeniero Juan Carlos Rojas expuso que los agricultores piuranos han identificado y evaluado aproximadamente ocho tipos de plagas que afectan el banano orgánico, entre las más frecuentes se tienen a:

- Picudo negro o gorgojo: este insecto es considerado mundialmente como el principal autor de pérdidas en las plantaciones de banano orgánico. Su presencia afecta principalmente al fruto de las plantas de banano, es decir, su reproducción le resta vitalidad a la planta, lo cual conduce a la caída de las plantas maduras.

Para controlar la reproducción de este insecto se deben realizar actividades como desmalezado y el deshoje (procesos descritos en el aparatado anterior), ya que, de no ser controlado, el daño puede ocasionar hasta un 50% de pérdidas para los agricultores.

- Sigatoka negra: esta enfermedad ataca directamente a las hojas del banano y se presenta en forma de rayas y manchas negras por debajo de dichas hojas. Esta presencia de manchas y rayas negras precipitan la muerte de la hoja. Se tienen seis escenarios que indican la presencia y evolución de esta enfermedad (Ing. Juan Carlos Rojas Llanque, 2012):

    1. Lesiones color amarillentas, no visibles a trasluz
    2. Rayas oscuras, color café entre 2 y 3 mm.
    3. Las rayas mencionadas en el punto dos se alargan, alcanzando incluso hasta los 2 o 3 cm de largo.
    4. Manchas negras en forma de elipse.
    5. En algunas ocasiones, aparición de manchas negras englobadas de un halo amarillento y centro semihundido.
    6. Características peculiares en la mancha tales como: centro desecado, color gris, halo amarillo brillante.

Para poder prevenir esta enfermedad los agricultores utilizan técnicas como control con productos orgánicos preventivos, siembra de variedades de fruta en dicha zona resistente a la plaga, buena limpieza y control diario.

- Sigatoka amarilla: es la enfermedad del banano que ataca las células de las hojas, se caracteriza por formar manchas de color amarillo en la hoja del banano y a diferencia de la Sigatoka negra esta enfermedad destruye la hoja del banano que interfiere en el crecimiento y calidad del racimo claramente.

De todas estas plagas, el trip de la mancha roja es la más temida por los productores, porque tiene la facilidad de esparcirse rápidamente entre las plantas y existen pocas alternativas para combatirla.

1.3.1 *Trip de la mancha roja*

En el año 2013, un estudio demostró que en el norte del país se ubican las principales áreas de producción de banano orgánico, dejando a Piura como el departamento con mayor producción con un 81%, mientras que, en Tumbes, la producción fue de 13%, Lambayeque 4% y La Libertad 2% (Juan Carlos Rojas Llanque, 2013).

El trip de la mancha roja es un insecto, el cual tiene como nombre científico *Chaetanaphothrips signipennis* (PIP Banano Orgánico), el cual suele absorber la sabia de la planta, hacer agujeros a las hojas para depositar sus huevos y reproducirse y además, causa manchas de color marrón o rojizo en las cáscaras de los bananos (Ver figura 14) e, incluso, causar la muerte de la planta (Juan Carlos Rojas Llanque, 2013).

El trip de la mancha roja se puede clasificar en dos grupos: cuando son ninfas (proceso de crecimiento) y cuando son adultas (proceso de reproducción), siendo las ninfas, las más dañinas para las plantas (Juan Carlos Rojas Llanque, 2013).

Figura 14. Banano atacado por el trip de la mancha roja
Fuente: Manejo integrado de banano orgánico (Juan Carlos Rojas Llanque, 2013) .

Cabe indicar que cuando el banano presenta manchas en sus cáscaras, este ya no puede ser comercializado ni exportado, ya que uno de los principales requerimientos de los mercados extranjeros es que el banano tenga un color verde uniforme, sin manchas ni daños.

1.3.2 *Manejo del trip*

En los últimos años, los productores piuranos han ideado un método para controlar y evitar la presencia de esta plaga en sus plantaciones, el cual realizan de forma interdiaria o semanal y anual, de acuerdo a sus recursos económicos.

Una forma de proteger a la planta es colocándole fundas o bolsas de protección; sin embargo, algunas veces cuando hay muchas plantaciones, la protección no se coloca a tiempo para evitar la propagación de la planta.

Otra forma de controlar la reproducción del trip de la mancha roja es limpiando la planta y realizando un corte temprano de cúcula. Algunos productores afirman que el trip de la mancha roja no resiste la exposición al sol, sino que prefiere lugares de sombra y las fundas de colores oscuros que se coloca a las plantas; por ello, una limpieza constante de las hojas secas que caen y/o de las flores a las que se les quitan los frutos, así como el uso de bolsas de colores verdes, azules o transparentes ayudan a prevenir la propagación de esta plaga.

Capítulo 2
Técnicas de aprendizaje supervisado

Al abordar la agricultura de precisión, es necesario relacionar valores arrojados por sensores con variables desconocidas, y de esta manera obtener una relación matemática entre variables dependientes e independientes. Para ello, se pueden recurrir a las técnicas de aprendizaje supervisado.

El aprendizaje supervisado es un conjunto de técnicas o estrategias que permite establecer una función a partir de los datos de entrada (datos de entrenamiento), la cual es capaz de predecir el valor deseado correspondiente a cualquier dato de entrada nuevo (Zhu, 2009).

Entre las técnicas más conocidas del aprendizaje supervisado, tenemos a la técnica de Support Vector Machine (SVM), el cual, a su vez se divide en dos tipos: algoritmos de regresión y algoritmos de clasificación., donde la diferencia entre ambos tipos es la salida, los algoritmos de regresión arrojan una salida como dato numérico y los algoritmos de clasificación arrojan una salida como dato categórico. Vapnik V. y Lerner (1963)

Las técnicas que implican el estudio de correlación entre variables, requieren de una buena base de datos, los cuales, deben ser estudiados y analizados previamente, para obtener resultados que demuestren un comportamiento similar a la vida real. Vapnik V. y Lerner (1963).

2.1 Normalización

La normalización de datos es una técnica de pre – procesamiento de datos, que tiene como finalidad que todas las variables de nuestro conjunto de datos estén a la misma escala. Por ejemplo, al suponer que se tiene una data con dos variables la primera corresponde al promedio de altura de personas de un determinado país en metros y la segunda el número de habitantes de ese mismo país (millones de habitantes), al comparar ambas variables, la primera variable numéricamente es muy pequeña comparada con la segunda variable, esto dificulta el entrenamiento del algoritmo, ya que una pequeña variación de los parámetros internos del algoritmo de proyección generará un gran cambio en la salida debido a la falta de normalización de sus variables.

Como se ha mencionado, la normalización es una técnica donde se requiere que todas las variables tengan la misma escala y, a la vez, puedan ser comparables entre ellas.

Por medio de la estandarización, se deduce:

- Todas las variables tienen un valor promedio igual a cero ($\bar{x} = 0$)
- Una desviación estándar igual a 1 ($\sigma = 1$)

Para lograr la estandarización se aplica:

$$z_i = \frac{x_i - \bar{x}}{\sigma} \quad (1)$$

Donde:

- $\bar{x}$ es el valor promedio de cada variable
- x_i es valor de cada variable
- σ es la desviación estándar

2.2 Regresión lineal

La regresión lineal es una técnica que permite encontrar una relación matemática o vínculo entre variables independientes y variables dependientes, donde las variables independientes tienen la posibilidad de cambiar, haciendo que las variables dependientes varíen automáticamente (Su, 2012).

El "cambio" de las variables independientes puede ser producido por factores cuantitativos o cualitativos. Un factor cuantitativo es aquel que puede medirse según niveles numéricos, tales como la temperatura, humedad, radiación solar, presión, entre otras; mientras que, los factores cualitativos son aquellos que no pueden medirse numéricamente.

Por ejemplo, si hablamos de una parcela de banano orgánico, se puede decir que:

- Factores cuantitativos: temperatura, humedad, presión, cantidad de veces regada al mes, cantidad de veces fumigada la parcela con productos orgánicos
- Factores cualitativos: calidad de la mano de obra en la cosecha de banano, sexo de las personas que cosechan, estado de la maquinaria (tinas, rieles)

Para determinar qué tipo de variables se pueden relacionar con la propagación del trip de la mancha roja en el banano orgánico, es necesario recordar que el crecimiento de esta plaga está ligado a factores como la temperatura, el nivel de humedad y la cantidad de riego de la parcela; siendo estos factores cuantitativos y es, por tanto, que la regresión lineal se trabajará con variables cuantitativas e independientes.

Asimismo, es necesario añadir que la regresión lineal se utiliza para predecir una respuesta cuantitativa Y sobre la base de una variable predictiva e independiente X. Por tanto, se puede afirmar que es el cálculo de la recta que mejor se comporta como respuesta a la

relación entre la respuesta y las variables que la explican, asumiendo que existe una relación entre ellas.

La representación gráfica de la regresión lineal es la de una recta que relaciona los valores del eje X con los valores del eje Y (ver Figura 15), donde:

- Eje X: numeración de la variable independiente
- Eje Y: numeración de la variable dependiente
- Los valores del eje Y son la respuesta de la variable que se está midiendo en el eje X.

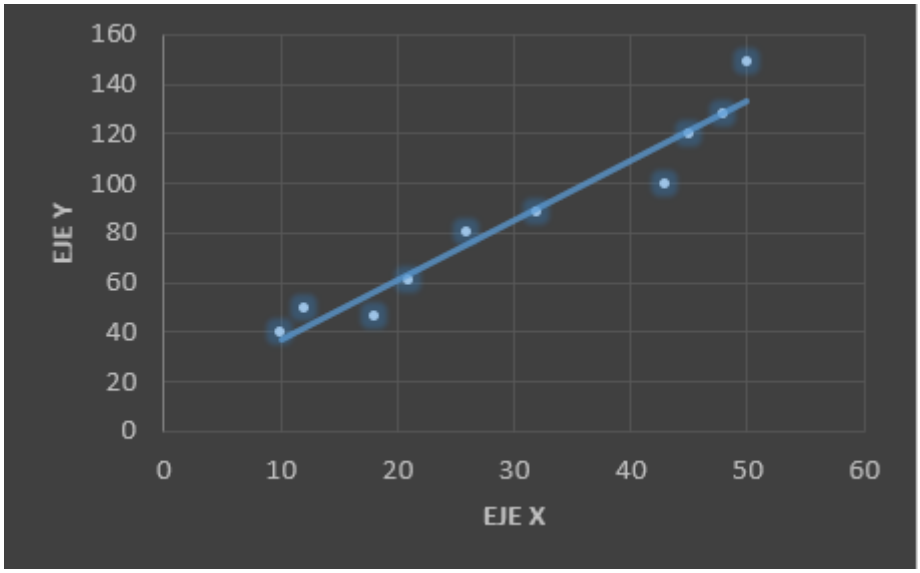

Figura 15. Representación gráfica de regresión lineal

2.2.1 *Tipos de Regresión lineal*

Los tipos de regresión lineal se relacionan con el número de variables, es decir, si se habla de una regresión lineal simple, se hace un enfoque de una única variable independiente (característica); sin embargo, esto no significa que dicha predicción sea la correcta. Además, existe también la regresión lineal múltiple, que posee más de 1 variable independiente (Su, 2012).

2.2.1.1 Regresión lineal simple. La regresión lineal simple tiene 1 sola variable independiente, que también se denomina variable predictora. Si recordamos la figura 16, podemos deducir que, la ecuación de dicha recta viene a ser definida de la siguiente manera:

$$y = b_0 + b_1 x \quad (1)$$

Donde:

- Y, es la variable dependiente
- x es la variable independiente o predictora
- b_0 es el término independiente constante que determina el valor de Y cuando X es cero

- b_1 es la pendiente de la recta, según su posición se puede concluir la variación o el cambio según los incrementos de los valores

Es sabido que, en la práctica, los valores de los coeficientes son desconocidos y que, además, las distancias entre los puntos y la línea de regresión son los residuos, donde se puede denominar valor observado, al valor del punto, y valor aproximado, al valor del punto en la recta (ver figura 16).

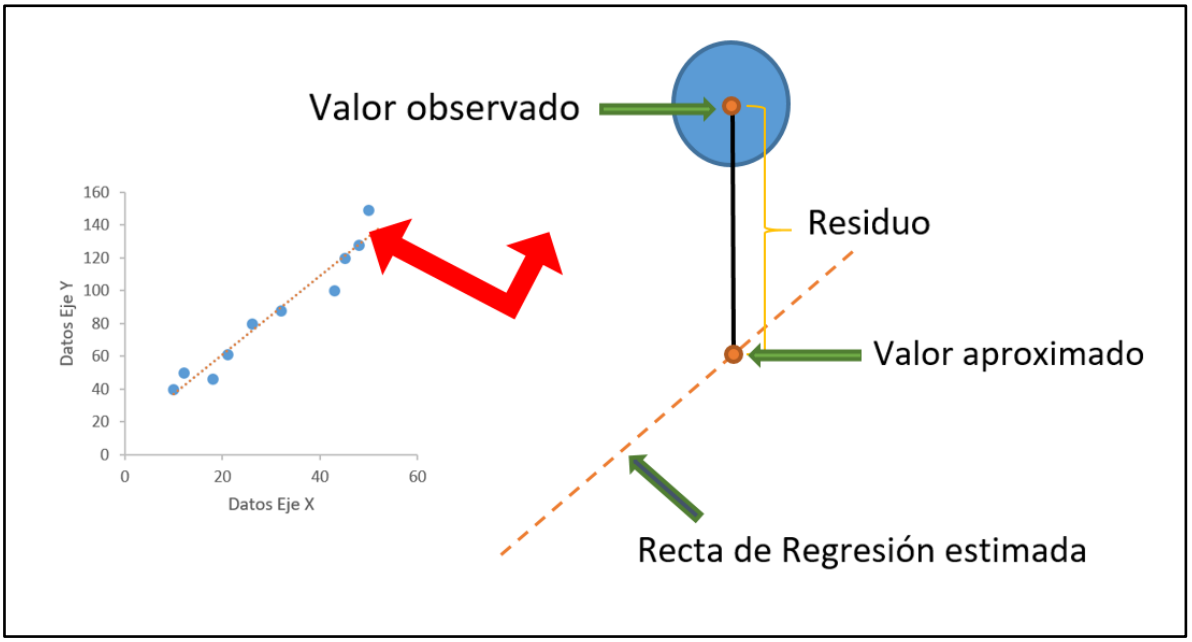

Figura 16. Representación del residuo

Para interpretar correctamente la regresión lineal representada en la figura 17, se debe observar que existen puntos más lejanos y más cercanos a la recta y que mientras más cercanos se encuentren, significa que existe un buen ajuste entre la recta de regresión y la variable independiente estudiada. Por otro lado, cuando los puntos están lejos de la recta, el residuo entre ambos ayudará a comprobar en la ecuación el buen ajuste de la misma.

Por tanto, la idea fundamental de la regresión lineal es tener una recta que se ajuste a los datos y que, a la vez, estos puntos no estén muy alejados de la recta; por ello, en la regresión lineal es frecuente el uso del método de los mínimos cuadrados (Su, 2012).

El método de los mínimos cuadrados es utilizado para minimizar los errores y determinar correctamente la recta. Este método consiste en determinar la recta que minimiza la suma de las distancias verticales cuadradas, desde los puntos hacia la recta. Existen 5 hipótesis del modelo de regresión lineal simple (Carrasquilla A, 2016):

a. Linealidad:

Indica que la relación lineal entre la variable independiente x y la variable dependiente y($f(x)$).

$$f(x) = \beta_0 + \beta_1 x \quad (2)$$

b. Homogeneidad

Indica que el error debe tener un valor promedio igual a cero (garantiza el buen ajuste de la recta con la variable).

$$E\,[\mathrm{u}i] = 0 \quad (3)$$

c. Homocedasticidad

Indica que la varianza de los errores es constante.

$$Var\,(\mathrm{u}i) = \sigma^2 \quad (4)$$

d. Independencia

Indica que los factores o características expuestas son las variables independientes.

$$E\,[\mathrm{u}i\ \mathrm{u}j] = 0 \quad (5)$$

e. Normalidad

Indica que los errores siguen una distribución normal.

$$\mathrm{u}i \sim N\,(0, \sigma) \quad (6)$$

Cuando se grafica una recta y se relaciona con la data, se espera, por lo general, que esta sea la adecuada para la predicción; sin embargo, surge la pregunta: ¿cómo asegurar que dicha recta o línea de regresión sea la adecuada o significativa? (Carrasquilla A, 2016) En respuesta, se han establecido las siguientes relaciones:

- Hipótesis nula

$$H_0\,, \beta_1 = 0 \quad (7)$$

- Hipótesis alternativa

$$H_0\,, \beta_1 \neq 0 \quad (8)$$

- P es la probabilidad y se encuentra entre 0 y 1

En un estudio realizado por (Carrasquilla A, 2016)se detalla que, para empezar un análisis, se debe verificar si el valor verdadero de la pendiente β_1 es igual a cero. Si la recta es totalmente horizontal, no se obtendrá pendiente o esta será igual a 0, por lo tanto, no existirá ninguna relación entre las variables; sin embargo, si la recta no es horizontal, significa que la

pendiente no es cero, y esto es un indicador de que existe una relación entre las variables estudiadas.

Posteriormente, se utiliza el parámetro α, el cual tiene el valor de 0.05 y ayudará a definir otra de las hipótesis mencionadas, ya que:

- Si, P > α, no se rechaza Ho
- Pero si P < α, entonces se rechaza Ho

Esto quiere decir que, se rechazaría la hipótesis nula, que afirma que la pendiente es igual a cero.

Después de determinar que realmente existe una relación entre las variables, y que esta a su vez es estadísticamente significativa, se procede a estudiar la variable de regresión; de esta forma, si esta variable característica independiente de regresión se relaciona en su mayoría con la variable dependiente, entonces explicaría la mayor parte de las variaciones de dicha variable dependiente (Carrasquilla A, 2016).

Para comprender la idea, es necesario realizar un gráfico de dispersión, que demuestra que los puntos están lo más próximos posibles a la recta, demostrando así que los residuos o el error es pequeño (ver figura 17).

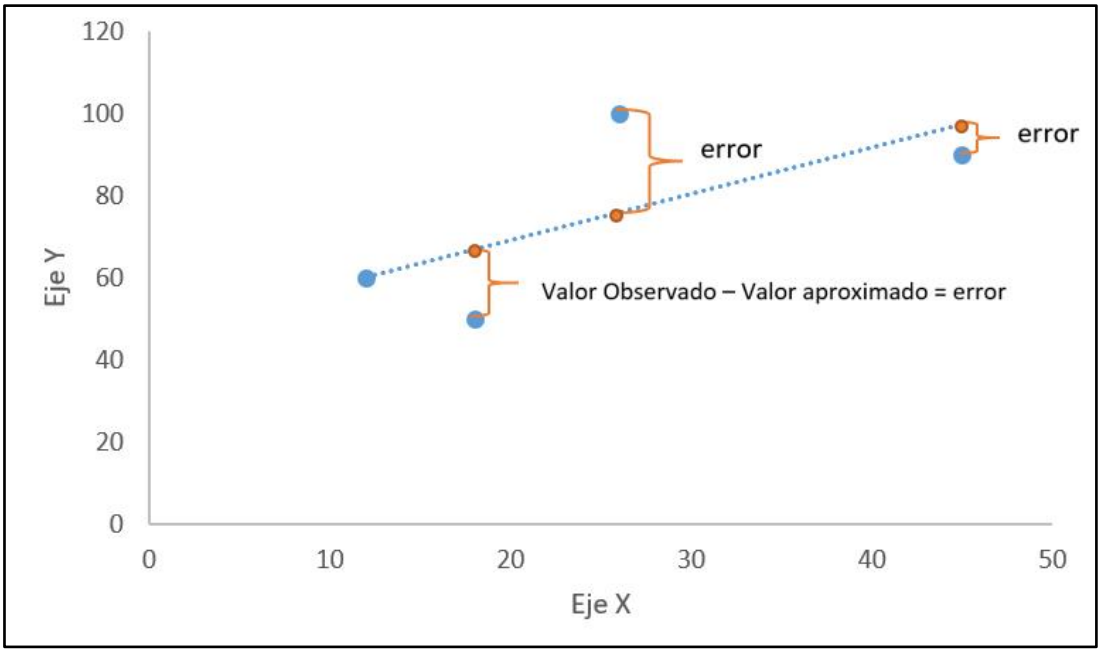

Figura 17. Gráfico de dispersión y error de regresión lineal

En la regresión lineal, la variabilidad permite reconocer cuán alejado está un resultado de la respuesta. Sus valores están en un rango de 0 y 1; sin embargo, suele expresarse en porcentaje para una interpretación más sencilla, entonces, el nuevo rango que tiene la variación será de 0% a 100%. El valor que se obtiene en la variabilidad indica si la recta de

regresión se alinea o se ajusta con la data, y a la vez, indica cuanta variación implica entre la variable independiente y dependiente (Carrasquilla A, 2016).

Otra forma de determinar que la variable independiente está relacionada con la variable dependiente es realizando una prueba de hipótesis de la pendiente. Como ya se explicó, la hipótesis nula confirma que la pendiente es cero; mientras que, la hipótesis alternativa especifica que la pendiente no es igual a cero.

Finalmente, para validar una regresión lineal, se deben cumplir tres condiciones:

- Los errores son independientes y aleatorios
- La distribución de los errores, es normal
- A lo largo de los valores de X, los errores tienen una varianza constante

2.2.1.2 Regresión lineal múltiple. El modelo de regresión lineal múltiple es similar al de regresión lineal simple, donde, además, se añaden más variables independientes dispuestas a tener alguna relación con las variables dependientes (Carrasquilla A, 2016). Se expresa con la siguiente ecuación:

$$y = f(x_1, \dots, x_k) + \in \tag{9}$$

Donde:

- y es la variable dependiente
- x es la variable independiente que explica alguna relación
- f es la ecuación del modelo de regresión
- $\in$ es el error independiente y aleatorio

Se traduce en la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \in \tag{10}$$

Según (Carrasquilla A, 2016) la regresión lineal múltiple se suele utilizar en dos ocasiones:

- Cuando el comportamiento de la variable dependiente Y, depende linealmente de las variables independientes X.
- Cuando un factor o variable no es suficiente para explicar el comportamiento de la variable Y.

Por ejemplo, cuando existen dos variables independientes que explican el comportamiento de una variable dependiente, la regresión lineal múltiple se expresa matemáticamente con la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \in \quad (11)$$

Al graficar el modelo, se obtiene la siguiente imagen:

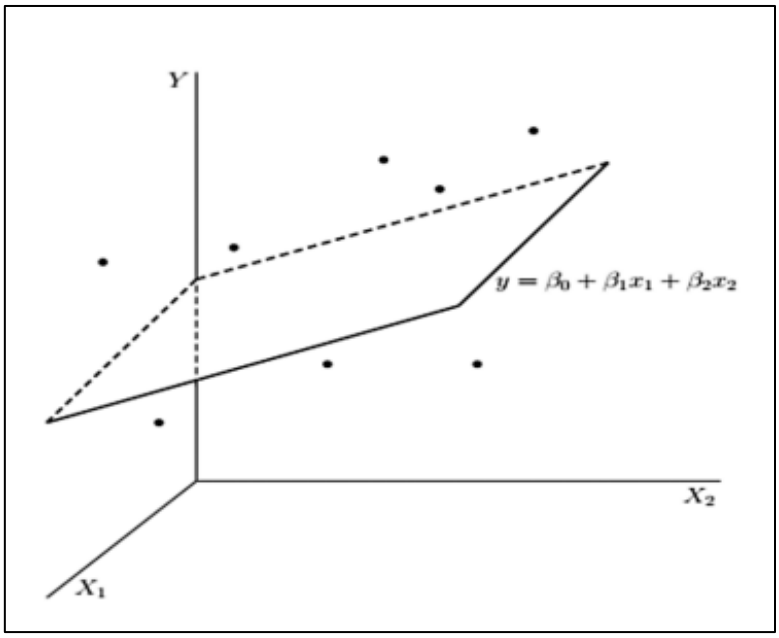

Figura 18. Modelo de regresión lineal con dos variables visto en 3D

Fuente: Tomado de "Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal" (Arys Carrasquilla - Batista, Alfonso Chacón - Rodriguez, Kattia Núñez - Montero, Olman Gómez - Espinoza, Johny Valverde, Maritza Guerrero - Barrantes, 2016)

De la misma forma que existen las hipótesis para la regresión lineal simple, también existen para la regresión lineal múltiple, los cuales son necesarias para verificar el ajuste de la recta. Una de las hipótesis afirma que, cuando se tiene los valores de las variables independientes, la ecuación es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \in_i \quad (12)$$

Donde:

- Los errores son independientes para las dos variables
- Las variables son, linealmente independientes
- El número de datos es mayor o igual a k+2

2.3 Support Vector Machine (SVM)

En la vida real, es muy baja o nula la probabilidad de encontrar una solución sencilla a un problema donde dos o más características encuentren relación mediante un ajuste de una recta de regresión lineal; por ello, resulta conveniente estudiar un algoritmo de aprendizaje supervisado de *Support Vector Machine (*SVM), el cual utiliza la técnica de clasificación. Vapnik V. y Lerner (1963).

El SVM es un algoritmo de aprendizaje supervisado dentro del campo de Machine Learning, y se utiliza para resolver problemas de clasificación de datos. Este algoritmo fue propuesto por Vapnik V. y Lerner (1963) y se enfoca en la separación de datos, que se realiza con la generación de un hiperplano optimizado (ver figura 21), también llamada función de decisión óptima, que separa los datos en dos "clases", previamente definidos (etiquetados).

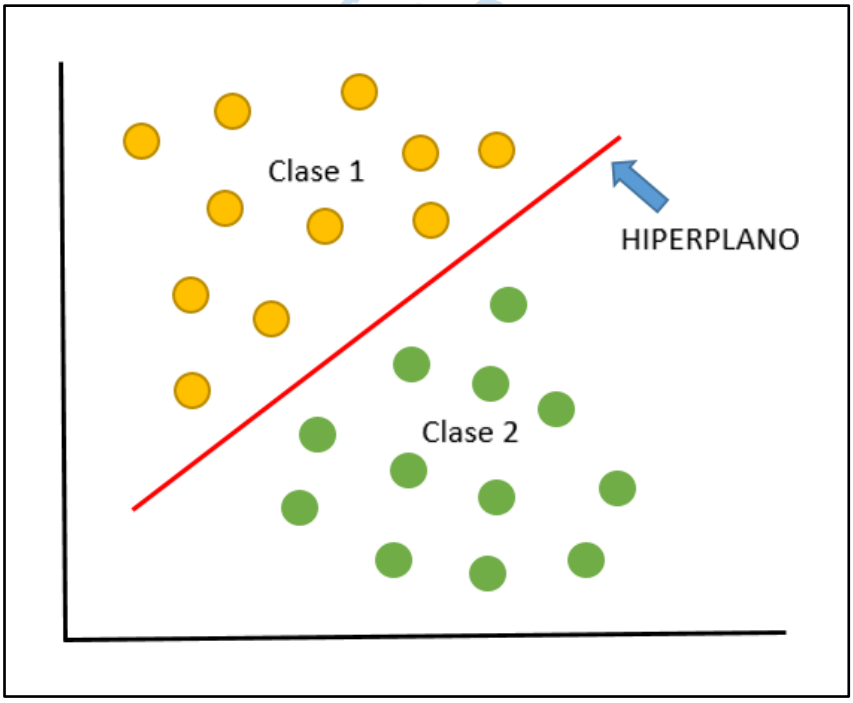

Figura 19. Separación de datos mediante un plano. Vista en 2D

El concepto de hiperplano optimizado hace referencia a la solución de un problema de optimización; es decir, es la mayor distancia (margen máximo) que existe entre los puntos más cercanos al hiperplano (vectores de soporte de cada una de las clases). Cabe mencionar, que para la aplicación de este algoritmo se utilizan datos de entrada linealmente separables.

Posteriormente, el SVM utilizó algoritmos no lineales para la resolución de problemas de clasificación de datos no separables. Asimismo, se implementó y adaptó para resolver problemas de regresión (Cortes, 1995).

2.3.1 *Formulación matemática*

A continuación, en la figura 20, se describe el esquema de la formulación matemática que sustenta al algoritmo SVM.

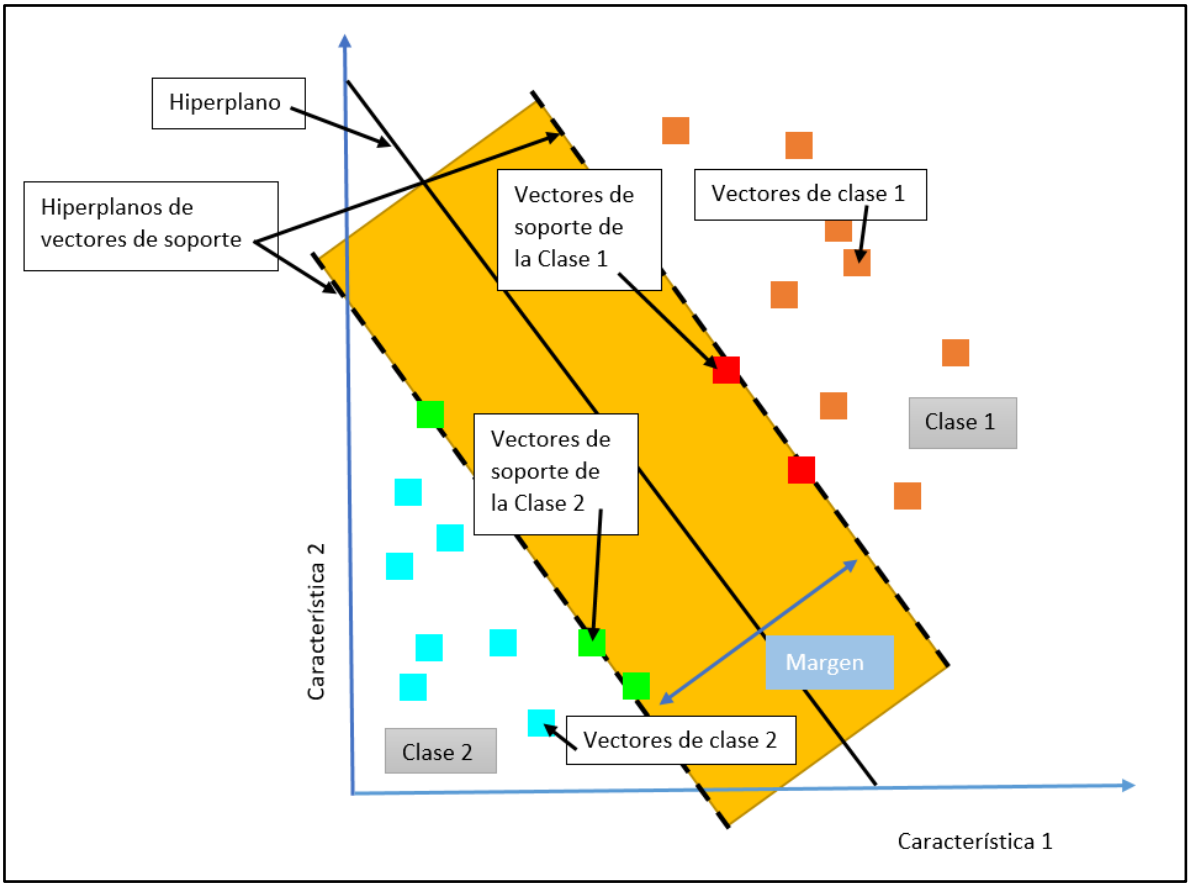

Figura 20. Esquema de la formulación matemática de SVM

Tal como se observa en la figura 20, los datos de entrada son los cuadrados, que poseen 2 características (característica 1 y característica 2). Para entender mejor el esquema, cada dato de entrada se considera un vector, el cual posee 2 dimensiones, por sus 2 características. Aquí, los datos de entrada (vectores) se han etiquetado y pertenecen a la clase 1 o a la clase 2, cada clase tiene vectores de soporte (que también son datos de entrada etiquetados), los cuales permitirán optimizar el hiperplano, maximizando el margen. Finalmente, se muestra el hiperplano de solución y los hiperplanos generados por los vectores de soporte.

En la figura 21, se observa los diferentes hiperplanos que se pueden generar para la separación de los datos de entrada (etiquetados). Si bien es cierto que, los 3 hiperplanos mostrados en la figura 21 separan los datos etiquetados correctamente, esto no quiere decir que son la mejor solución, ya que, el margen no es máximo. Esto quiere decir, que el margen de estos 3 hiperplanos es más estrecho que el hiperplano solución (optimizado).

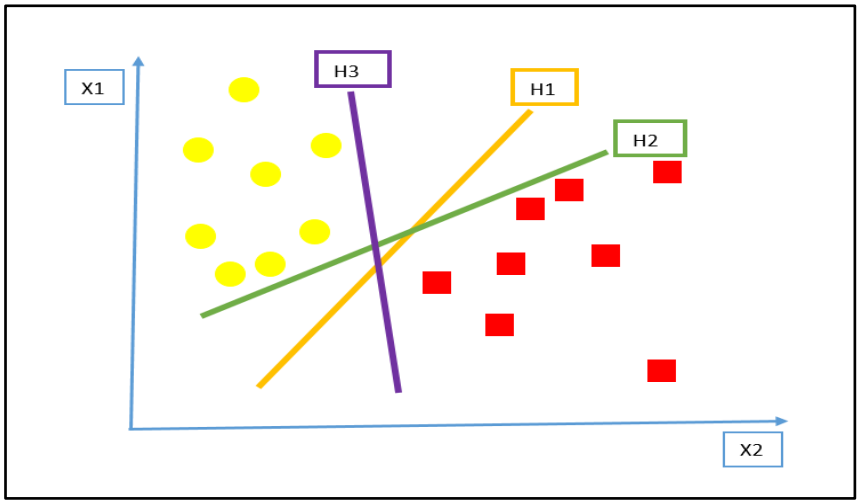

Figura 21. Hiperplanos que se pueden generar para la separación de datos de entrada etiquetadas

Para generar este hiperplano optimizado H, el hiperplano en cuestión debe cumplir con un margen de separación máximo, tal como se muestra en la figura 22.

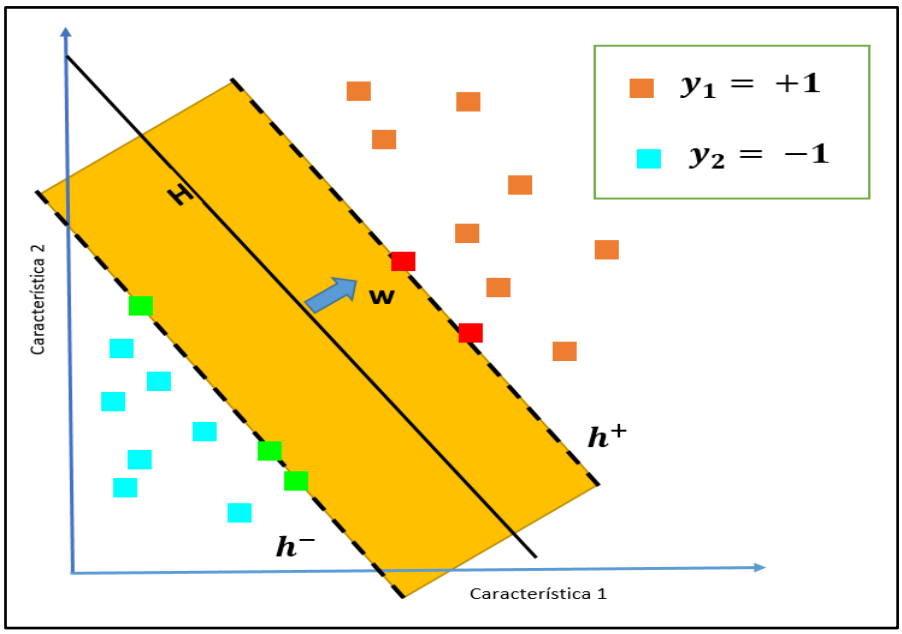

Figura 22. Ejemplo de hiperplano optimizado con margen y los vectores de soporte

Como existe un hiperplano optimizado H, cuya solución se puede expresar de la siguiente manera: a partir de un vector w, ortogonal al hiperplano optimizado H, y un coeficiente de intercepción b, tenemos:

Hiperplano optimizado H:

$$w^T * x_i + \mathrm{b} = 0 \qquad (10)$$

Este hiperplano solución es un hiperplano medio entre otros 2 hiperplanos (ver figura 24): h^+ y h^-, que contiene los vectores de soporte de cada clase respectivamente. Utilizando como base la ecuación 10 y reemplazando, tenemos:

- Hiperplano h^+:

$$w^T * x_i + \mathrm{b} = 1 \tag{11}$$

- Hiperplano h^-:

$$w^T * x_i + \mathrm{b} = -1 \tag{12}$$

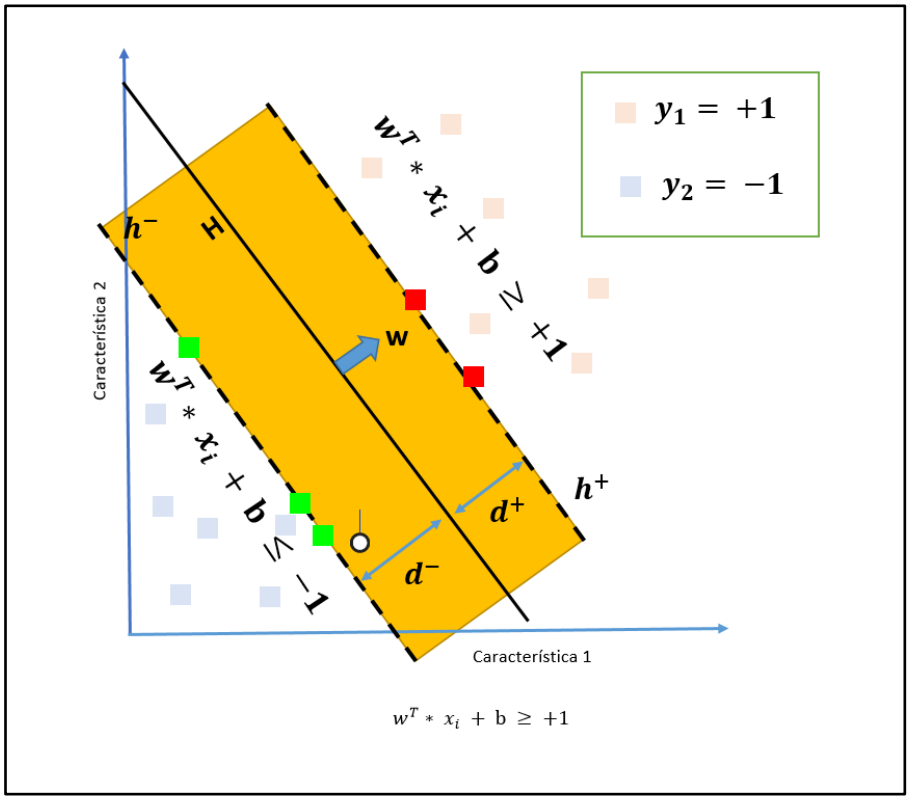

Figura 23. Hiperplano optimizado y condición de clasificación

Por lo tanto, a partir de las ecuaciones 11 y 12, tenemos dos condiciones de clasificación:

$$w^T * x_i + \mathrm{b} \geq +1 \tag{13}$$

$$w^T * x_i + \mathrm{b} \leq -1 \tag{14}$$

Entonces, cuando se aplica las condiciones de clasificación de los parámetros del hiperplano optimizado H a las muestras positivas, debe dar un valor mayor a 1, y cuando apliquemos a las muestras negativas, debe dar un valor menor de -1.

Estas 2 condiciones se pueden simplificar al condicional general de la siguiente manera:

$$y_i\ (w^T * x_i + \mathrm{b}) \geq +1 \tag{15}$$

Donde, la variable y_i es el valor de la etiqueta de cada clase, siendo +1 las muestras positivas y -1 las muestras negativas.

Por su parte, el margen se puede calcular midiendo la distancia entre los 2 hiperplanos h^+ y h^- , de la siguiente manera:

$$d^+ = d^- = \frac{|wx + b|}{\|w\|} = \frac{1}{\|w\|} \tag{16}$$

Entonces, utilizando la ecuación 16, tenemos:

$$margen = d^+ + d^- = 2\frac{1}{\|w\|} \tag{17}$$

Se puede observar que el margen solo depende de los parámetros w del hiperplano.

Maximizar el margen equivale a minimizar el problema inverso, por tanto, se tiene:

$$Minimizar\ \Phi(w) = \frac{1}{2}\|w\|\text{^}2 \tag{18}$$

Además, a esta minimización se le añade la condición de clasificación (ecuación 15), lo que ocasionaría que el problema de clasificación se convierta en un problema de optimización cuadrática.

2.3.2 *Optimización cuadrática*

Acorde a la formulación matemática, el problema de clasificación de SVM se puede transformar en un problema de optimización cuadrática. Para ello, es necesario recordar que el objetivo del problema de clasificación de SVM es obtener los parámetros w y b del hiperplano solución (ecuación 10). Vapnik V. y Lerner (1963).

Para resolver el problema, se debe plantear como un problema de optimización cuadrática, que consiste en la minimización de la función Φ(w) que está sujeta a la restricción “condición de clasificación” (ecuación 15). Además, requiere de una función auxiliar llamada Lagrangiano L, que se plantea como la suma de la función que se quiere optimizar más las restricciones a la que está sujeto el problema (condición de clasificación – ecuación 15), y a su vez, estas restricciones se multiplican por un factor α, que se le conoce como multiplicadores de LaGrange.

Se formula el Lagrangiano L como:

$$L\ (x, \alpha\ ) = f(x) + \sum_i \alpha_i \tag{19}$$

$$L\ (x, \alpha\ ) = f(x) + \sum_i \alpha_i * g_i(x) \ldots \forall\ \alpha_i \geq 0 \tag{20}$$

De él, se deduce que la función a optimizar es:

$$F(x)\ \rightarrow Q(w) \tag{21}$$

Además, se considera las condiciones de clasificación:

$$G(x)\ \rightarrow y_i\ (w^T * \ x_i\ +\ \mathrm{b}) \geq\ 1 \tag{22}$$

Y α → Multiplicadores de LaGrange / factores

Entonces, el problema optimización cuadrática se puede resolver de la siguiente manera (considerar ecuación 18 y ecuación 15, respectivamente):

$$\left.\begin{array}{l} Minimizar\ \Phi(w) = \frac{1}{2}\|w\|\text{^}2 \\ Sujeto\ a{:}\ y_i\ (w^T * \ x_i\ +\ \mathrm{b}) \geq\ +1 \end{array}\right\} \quad max_\alpha(min_{w,b}\big(L\ (w, b, \alpha\ )\big))$$

A esta optimización se le conoce como **SVM PRIMAL.**

El Lagrangiano L del problema de SVM es:

$$L\ (w, b, \alpha\ ) = \frac{1}{2}\ w^T * w - \sum_i \alpha_i (y_i\ (w^T * \ x_i\ +\ b) - 1$$

Para minimizar el Lagrangiano L, el cual está representado por min w, b (L (w, b, α)), se necesita derivar parcialmente L respecto a w y b, obteniendo:

$$\frac{dL}{dw} = w - \sum_i \alpha_i * y_i * x_i = 0 \quad \Rightarrow \quad w = \sum_i \alpha_i * y_i * x_i \tag{23}$$

$$\frac{dL}{db} = -\sum_i \alpha_i * y_i = 0 \quad \Rightarrow \quad \sum_i \alpha_i * y_i = 0 \quad \alpha_i \geq 0 \tag{24}$$

Como se puede apreciar en las ecuaciones 23 y 24, el valor óptimo del parámetro w depende de la combinación lineal de x_i e y_i (muestras/valores de entradas y etiquetas respectivas).

Para incluir la solución del parámetro w en el Langrangiano del SVM, es necesario suponer que estamos en condiciones de convexidad, obteniéndose esta nueva formulación:

$$L\,(w,b,\alpha\,) = \frac{1}{2}\left(\sum_i \alpha_i y_i\, x_i\right)^T \left(\sum_j \alpha_j y_j\, x_j\right) - \sum_i \alpha_i y_i \left(\sum_j \alpha_j y_j\, x_j\right)^T x_i + \sum_i \alpha_i \quad (25)$$

$$L\,(w,b,\alpha\,) = -\frac{1}{2}\sum_i\sum_j \alpha_i * \alpha_j * y_i * y_i * x_i^T * x_j + \sum_i \alpha_i \quad (26)$$

Como se observa, el Lagrangiano L de SVM ya no depende del parámetro w, sino de las muestras (valores de entrada y sus etiquetas) y el factor α (multiplicadores de LaGrange). Así, se reformula la función L en una función O (α):

$$O\,(\alpha) = -\frac{1}{2}\sum_i\sum_j \alpha_i * \alpha_j * y_i * y_i * x_i^T * x_j + \sum_i \alpha_i \quad (27)$$

Entonces, se procede a maximizar la función O (α) que está sujeta a la siguiente restricción, la cual se obtuvo de la derivada parcial de L respecto a α (ecuación 24).

Recordando restricción ecuación 24: $\sum_i \alpha_i * y_i = 0 \qquad \alpha_i \geq 0$

Queda de la siguiente manera:

$$Maximizar\ O\,(\alpha) = -\frac{1}{2}\sum_i\sum_j \alpha_i * \alpha_j * y_i * y_i * x_i^T * x_j + \sum_i \alpha_i \quad (28)$$

$$Sujeto\ a: \sum_i \alpha_i * y_i = 0 \qquad \alpha_i \geq 0$$

Finalmente, se obtiene un nuevo problema de optimización cuya solución es la solución del SVM (que se le conoce como SVM DUAL). Se analiza la formulación del SVM dual:

$$w = \sum_i \alpha_i * y_i * x_i \quad (29)$$

Se puede observar que la solución óptima del parámetro w, depende la combinación lineal de las muestras de entrenamiento (valores de entrada y sus respectivas etiquetas), pero no de todas las muestras. Esto se debe a una parte de la restricción donde α ≥ 0, lo que significa que algunos valores de los multiplicadores de LaGrange son iguales a 0.

Por lo tanto, la solución óptima del parámetro w, será resultado de la combinación lineal de aquellas muestras que tenga asociado un factor α (multiplicador de LaGrange) mayor que 0; a estas muestras se les conoce como vectores de soporte. Estos vectores de soporte se ubican en h1 o h2, de tal manera que cumple con las condiciones de la ecuación 15 $(y_i\ (w^T * x_i + \mathrm{b}) \geq +1)$**.** Además, los vectores de soporte son los más cercanos al hiperplano optimizado.

2.3.3 *Condiciones de KKT*

Para encontrar el valor de b, se deben emplear las condiciones de Karush, Kuhn y Tucker (KKT), las cuales sirven para llegar a una solución óptima de un problema de optimización general. Vapnik V. y Lerner (1963).

Al revisar la teoría de las condiciones de KKT, es posible plantear el siguiente problema de optimización (minimizar la función f(w) que tiene la restricción gi(w)):

Min f(w)

s.t. (restricciones) $g_i\ (w)\ \leq 0, i = 1, \dots, k$

Se define una nueva función lagrangiana:

$$L\ (w, \beta\ ) = f(w) + \sum_{i=1}^{k} \alpha_i * g_i(w) \tag{30}$$

Donde:

- α son multiplicadores de LaGrange
- f es una función convexa, α i >=0.

Entonces, como se quiere garantizar la existencia del mínimo, para que la función se pueda minimizar, se deben presentar las siguientes condiciones de KKT:

- dL(w*, α*)/dw = 0
- dL(w*, α*)/dα*= 0
- (α*i) *gi(w*) =0, i=1, …,k (Condición complementaria de KTT)
- gi(w*) =< 0, i=1, …,k
- α*i>= 0 i=1, …, k

Por lo tanto, si se toma como base la forma primal de SVM, es decir, si se toma su lagrangiano y la condición complementaria de KKT, se puede obtener el valor de b^* y w^*:

$$\mathrm{w} * = \sum_{i=1}^{l} y_i * \alpha_i *\ x_i \tag{31}$$

$$b^* = y_i - w^* * x_i = y_i - \sum_{i=1}^{l} \alpha_i^* * y_i * x_i * x_j \quad (32)$$

Finalmente, el hiperplano óptimo es:

$$f(x, b^*, w^*) = \sum_{i \in vs} y_i * \alpha_i * x_i \cdot x + b^* \quad (33)$$

2.3.4 *Función de Kernel*

Existen casos donde los datos no son linealmente separables (figura 24) y, para ellos, se utiliza la función kernel, que tiene como objetivo incrementar las dimensiones de los vectores de entrada X. En una transformación no lineal O(x), este incremento de dimensiones tiene a su vez como objetivo que la data sea linealmente separable. Vapnik V. y Lerner (1963).

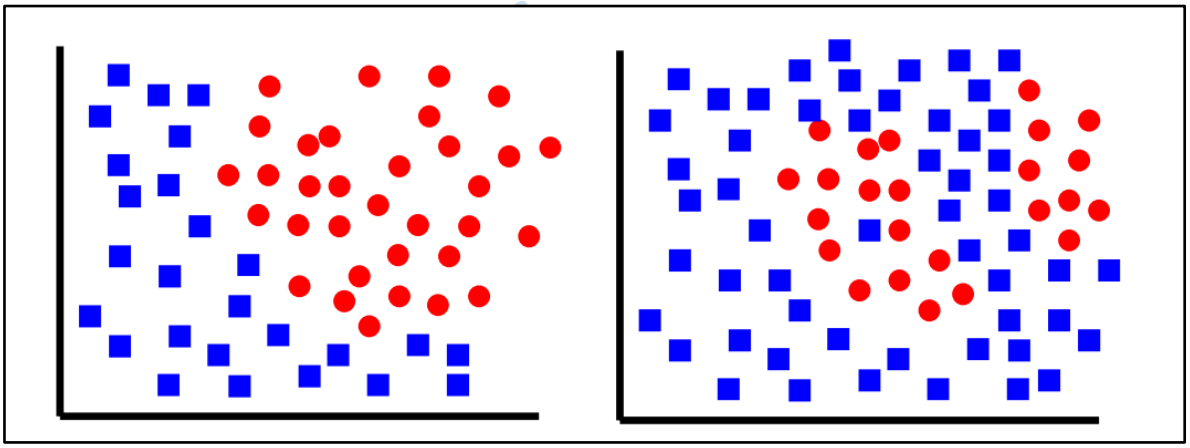

Figura 24. Ejemplos donde un separador lineal no sería eficiente

A continuación, en la figura 25 se presenta un ejemplo gráfico de la función kernel.

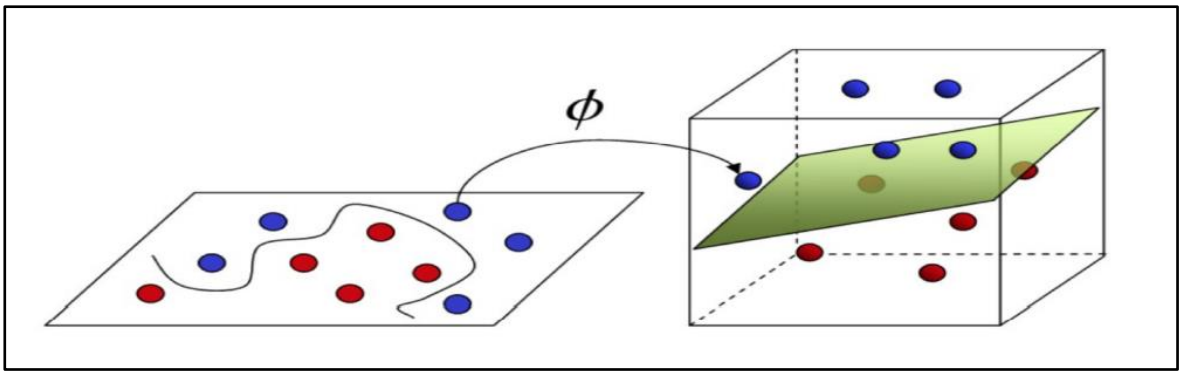

Figura 25. Transformación no lineal de la función kernel
Fuente: Aplicación de support vector machine al mercado colombiano, (Castañeda, 2019)

De la gráfica, se puede deducir que la función O: X-> F, es la función que hace que a cada de entrada X (n dimensiones) le corresponda un punto de salida en F (m dimensiones), siendo m>n.

Lo fundamental de esta estrategia es encontrar ese hiperplano óptimo que separa los datos transformados en 2 clases. Por lo tanto, si al problema de optimización se le aplica la función kernel O, quedaría de la siguiente manera:

$$\max \sum_{i=1}^{n} \alpha_1 - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_j \lambda_i y_i y_j \langle \Phi(x_i), \Phi(x_j) \rangle \tag{34}$$

El mismo que está sujeto a la siguiente restricción: $\sum_{i=1}^{n} \lambda_i y_i = 0$

2.4 Twin Support Vector Machine (TWSVM)

El *Twin Support Vector Machine* (TWSVM) fue propuesto por primera vez en 2007 por el investigador Jayadeva en su investigación llamada *Twin Support Vector Machines for Pattern Clasification.* Este método se utiliza para la separación o clasificación de datos linealmente separables en dos clases, basada en la construcción de dos hiperplanos no paralelos (un hiperplano positivo y otro negativo) para cada una de las clases del total de las muestras (Jayadeva, 2007), tal como se muestra en la figura 26.

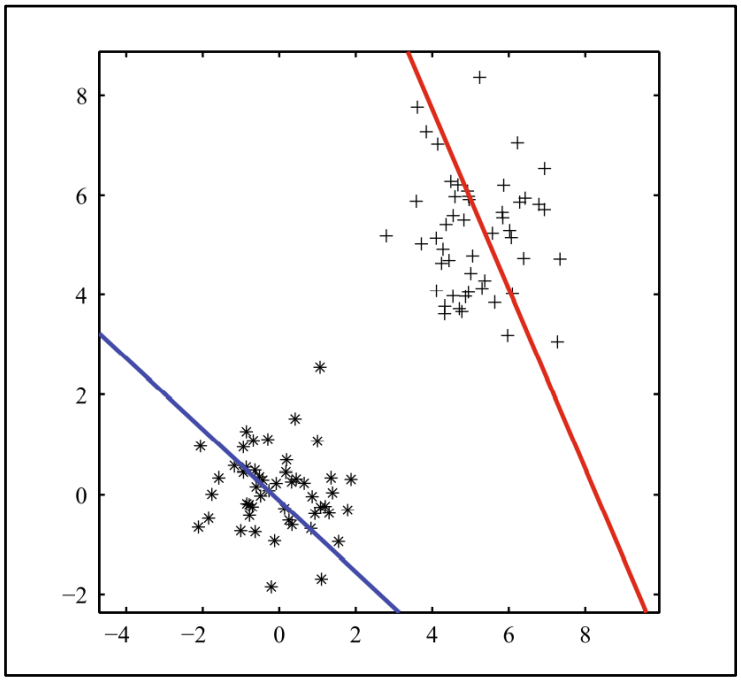

Figura 26. Ilustración geométrica de TWSVM
Fuente: *Twin Support Vector Machines for Pattern Clasification,* (Jayadeva, 2007)

Lo fundamental en el TWSVM es que cada hiperplano de una clase esté lo más lejos posible de los datos de la otra clase y más cerca de los datos de su misma clase. Vapnik V. y Lerner (1963) .Si existen nuevas muestras (nuevos datos de entrada), se les asigna una determinada clase y para ello, se mide la distancia desde la nueva muestra hasta cada uno de los 2 hiperplanos (h^+ y h^-). A partir de allí, se selecciona la menor distancia o el hiperplano

más cercano a la nueva muestra, haciendo que esta muestra pertenezca a la clase ligada a ese hiperplano en cuestión.

La diferencia más resaltante entre el SVM y el TWSVM es que en el primero se debe resolver un solo problema grande de programación cuadrática (QPP), mientras que en TWSVM se pueden resolver 2 problemas pequeños de programación cuadrática (QPP). Esto último se debe a la división de la data total en 2 (ver figura 27); por lo tanto, los patrones que existan para la clase 1 se deben tener en cuenta al momento de resolver el problema de programación cuadrática de la otra clase y viceversa. Por consiguiente, el TWSVM resulta ser más eficiente que el SVM (le toma menos tiempo resolver esos 2 problemas de programación cuadrática QPP).

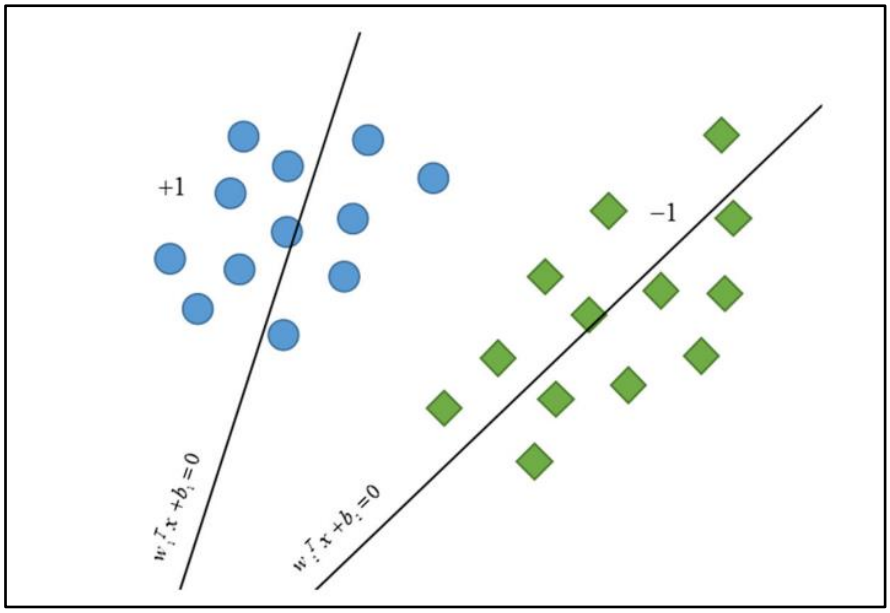

Figura 27. Interpretación Geométrica de TWSVM, con formulación matemática de los 2 hiperplanos

2.4.1 *Soft Margin (Margen blando)*

Antes de entrar a la formulación matemática del TSVM, es necesario comprender qué es el margen blando, para que sirve y cómo se implementa en SVM.

Tal como lo explica autor Vapnik V. y Lerner (1963), el margen blando se aplica cuando algunas muestras (datos de entrada) no son linealmente separables, tal como se observa en la figura 29, y por tanto, no existe un hiperplano que pueda separar las 2 clases. En este caso, se puede permitir que algunas muestras (puntos en la gráfica) se clasifiquen erróneamente para generar un hiperplano que tenga un margen lo más grande posible y un error de clasificación lo más pequeño posible.

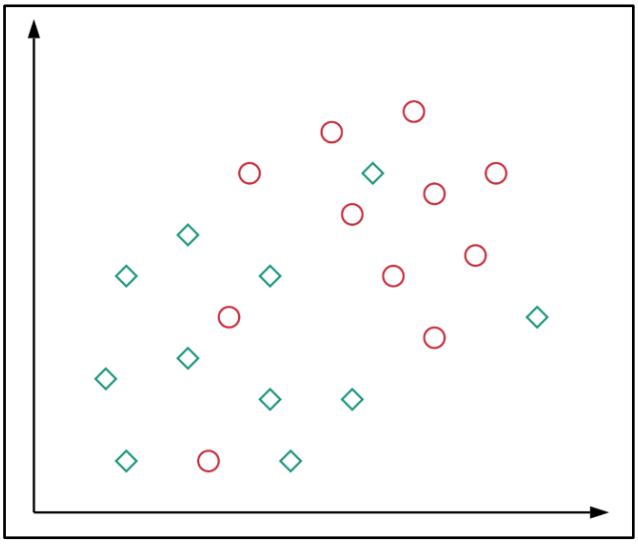

Figura 28. Muestras no linealmente separables

Por lo tanto, la idea detrás de la aplicación del margen blando es permitir que el SVM cometa un cierto número de errores de clasificación, pero manteniendo un margen grande, con lo cual muchas muestras (puntos) se puedan clasificar correctamente. En la figura 29, se presenta un ejemplo con todas las muestras posibles.

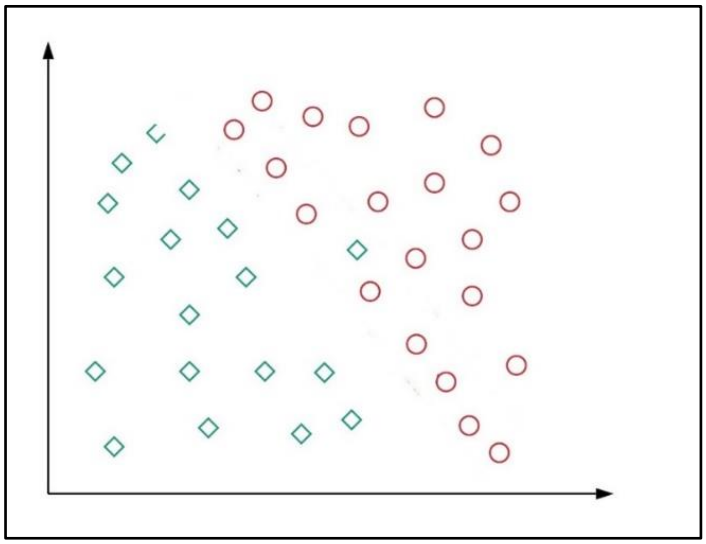

Figura 29. Total de muestras posibles

Al tomar unas muestras al azar para desarrollar un SVM que permita clasificar las muestras seleccionas en 2 clases (Ver figura 30), se puede apreciar que el hiperplano de color rojo separa perfectamente las 2 clases de las muestras, pero la función principal del SVM es predecir a que clase pertenece cada muestra nueva, que no hayan participado en la creación del hiperplano (rojo).

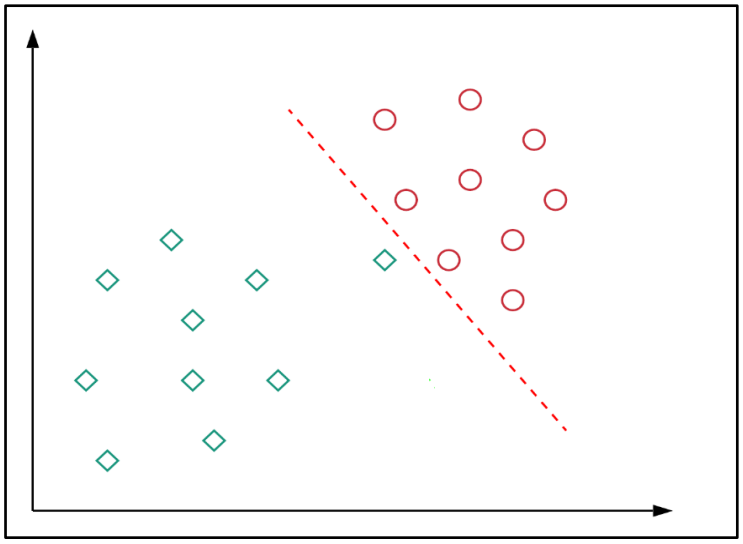

Figura 30. Muestras seleccionadas al azar con un hiperplano rojo, que no admite errores - Hard Margin

Si se insertan nuevas muestras al clasificador, como en la figura 31, se puede apreciar que existen muchos errores de clasificación (círculos verdes), por lo que si forzamos que el SVM tenga un margen "duro" (hard margin), es posible que al momento de predecir la clasificación de muestras nuevas, el error de predicción se incremente considerablemente.

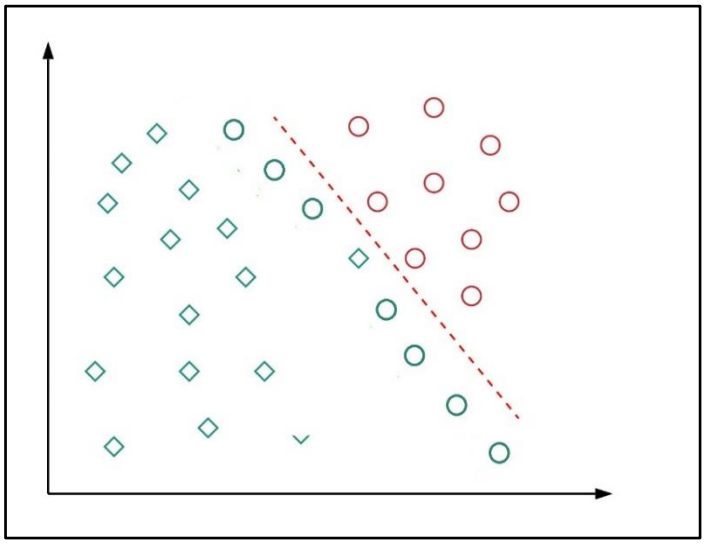

Figura 31. Círculos verdes mal clasificados

Lo contrario ocurre si es que se selecciona un hiperplano con un margen amplio (hiperplano verde), tal como se muestra en la figura 32.

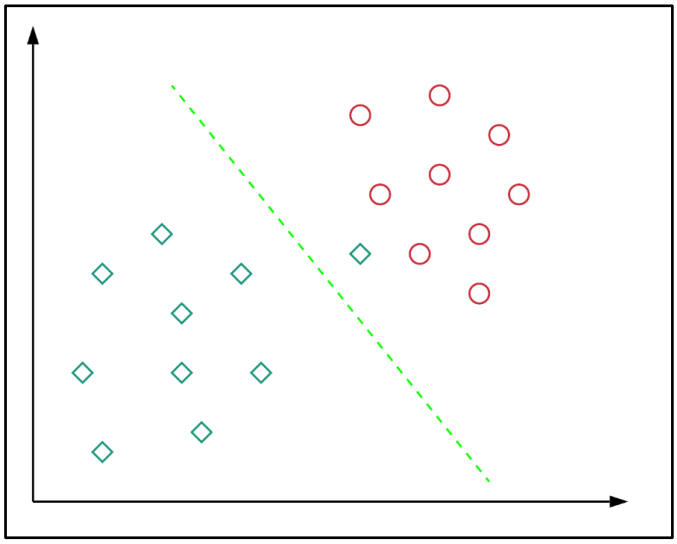

Figura 32. Muestras seleccionadas al azar con un hiperplano verde, que admite un error (cuadrado verde - Soft Margin)

Cabe indicar que, al momento de introducir nuevas muestras, se disminuye el error de predicción, de esta forma, en la figura anterior, los círculos verdes estarían bien clasificados, aunque existiría un error (cuadrado verde clasificado como círculo rojo), tal como se aprecia en la figura 33.

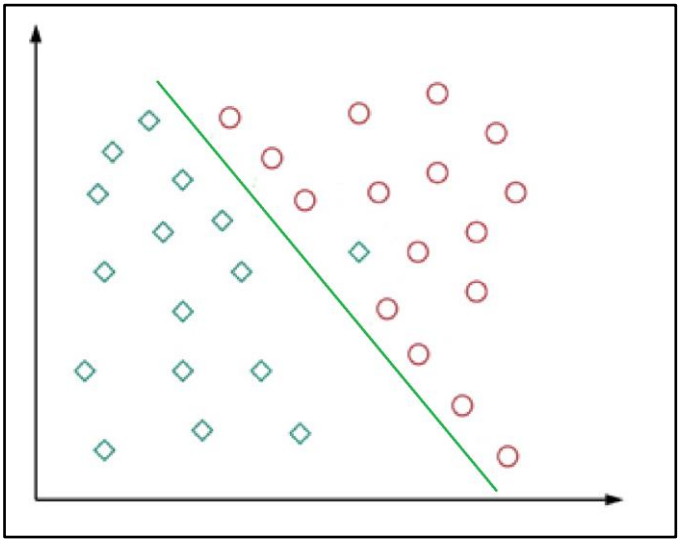

Figura 33. El error de clasificación de muestras nuevas disminuye considerablemente

Con este ejemplo, se puede concluir que al admitir errores al momento de implementar el hiperplano óptimo de separación del SVM, se incrementa el poder de predicción para muestras nuevas.

2.4.1.1 Formulación matemática. En la formulación matemática del Soft Margin en SVM se debe tener en cuenta que todos los errores no son iguales: hay errores más grandes y otros más pequeños. De esta manera, los datos de entrada que se encuentran en el lado equivocado y están más alejados del hiperplano que corresponde, representan un error mayor por lo que deben tener una mayor penalización (Ver figura 36).

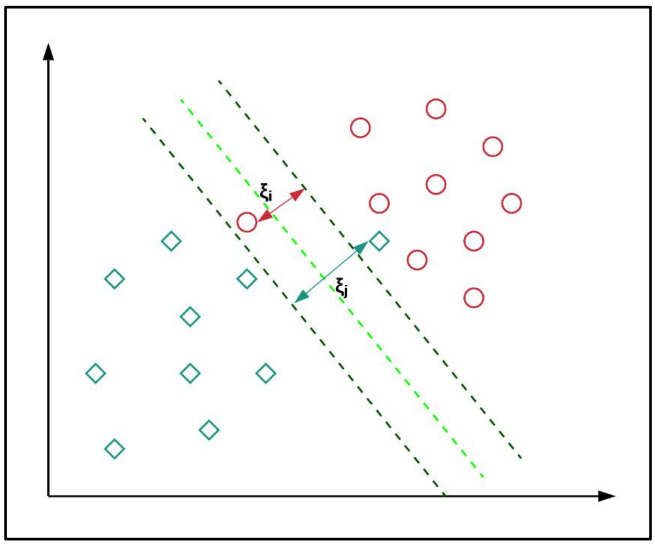

Figura 34. Personalización de los datos de entrada en el lado equivocado

Al plantear la formulación matemática, se deben tener en cuenta las condiciones generales para la clasificación en SVM con margen duro, que se expresan en la siguiente ecuación:

$$y_i\,(w^T * x_i + b) \geq 1 \qquad (35)$$

Entonces, para definir las condiciones generales de clasificación de SVM para un margen suave, se debe considerar el error dentro de la ecuación 35, de la siguiente manera:

$$y_i\,(w^T * x_i + b) \geq 1 - E \qquad (36)$$

Donde:

$$E \geq 0$$

En la condición planteada, se puede observar que existe una nueva restricción (E), que se interpreta como una variable de holgura: el valor de E será la distancia del margen de la clase correspondiente, siempre y cuando esté en el lado equivocado del margen, en caso contrario, tendrá el valor de 0.

Por lo tanto, los puntos más lejos del margen en el lado equivocado tendrán más penalización. (1 – E) también puede ser interpretado como la confianza o precisión de la clasificación, donde:

- Si E=0, se entiende la muestra o dato de entrada ha sido clasificada correctamente.
- Si E>0 significa que el clasificador ha cometido un error y por tanto, existe una penalización de E.

Para el SVM de margen duro, se debe resolver el siguiente problema de programación cuadrática:

$$Minimizar\ \Phi\ (w)_{w.b} = \frac{1}{2}\ \|w\|^2 \tag{37}$$

El problema está sujeto a:

$$y_i(w^T x_i + b) \geq 1 \tag{38}$$

Al implementar el *soft margin*, se obtiene lo siguiente:

$$\boldsymbol{Minimizar\ O(w) = \frac{1}{2}\ \|w\|^2 + C^*(suma\ de\ errores)} \tag{39}$$

La restricción se expresa con la siguiente ecuación:

$$y_i\ (w^T * x_i + b) \geq 1 - E, donde\ E \geq 0 \tag{40}$$

Dando como resultado:

$$\min \qquad \frac{1}{2} w^T + C \sum_{i=1}^{m} \xi_i \tag{41}$$

$$s.t. \quad y_i\ (x_i^T + b) \geq 1 - \xi_i, donde\ \xi_i \geq 0 \tag{42}$$

Se puede apreciar que existe una nueva variable C, un hiperparámetro que tiene la función de maximizar el margen y minimizar los errores. Cuando el valor de C es pequeño, no reciben mucha importancia los errores de clasificación; en cambio, cuando el valor de C es grande, el margen se vuelve más pequeño porque da mucha importancia a los errores de clasificación, para evitar una clasificación errónea.

Aplicando el Lagrangiano al problema de programación cuadrática, resulta la siguiente ecuación:

$$\int (w, \xi, b, \alpha) = \frac{1}{2}\ w^T w + C \sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i\ [1 - \xi_i - y_i\ (x_i^T + b)] \tag{43}$$

Por tanto, para encontrar la formal dual del problema de SVM, se debe minimizar el Langrangiano $\int (w, \xi, b, \alpha)$ respecto a w, ξ, b, dando como resultado:

$$D \begin{cases} max_{\alpha}\, \theta(\alpha) = \sum_{i=1}^{m} \alpha_1 - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_j \alpha_i \langle \Phi\left(x^{(i)}\right), \Phi\left(x^{(j)}\right) \rangle \\ 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{m} y_i * \alpha_i = 0 \end{cases}$$

Donde:

**< O(x), O(x) >* es la función kernel de las muestras (datos de entrada). Asimismo, al usar la derivada del Lagrangiano respecto a w, se obtiene:

$$\frac{dL}{dw} = w - \sum_{i} \alpha_i y_i x_i \tag{44}$$

Donde:

$$w = \sum_{i} \alpha_i y_i x_i = 0 \tag{45}$$

Al reemplazarlo, se obtiene:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i^T \tag{46}$$

Finalmente, se utilizan las condiciones de complementariedad de KKT, dando como resultado:

$$\alpha_i = 0 \qquad \Rightarrow y_i \left[x_i^T * w + b\right] \geq 1 \tag{47}$$

$$\alpha_i = C \qquad \Rightarrow y_i \left[x_i^T * w + b\right] \leq 1 \tag{48}$$

$$0 \prec \alpha_i \prec C \quad \Rightarrow y_i \left[x_i^T * w + b\right] = 1 \tag{49}$$

Es necesario recordar que, α será distinto a cero solo para los vectores de soporte y que el conjunto de vectores de soporte ahora incluye todos los datos en el límite del margen, así como aquellas muestras que están en el lado incorrecto del margen.

2.4.2 *Formulación matemática de TWINSVM*

Al conjunto de datos, linealmente separables en 2 clases, se le llama data de entrenamiento y tiene m ejemplos. Por ejemplo, en la siguiente fórmula:

$$T_C = \{(x^{(i)}, y_i), x^{(i)} \in \mathbb{R}^n,\ y_i \in \{-1, +1\}, (i = 1, 2, \dots, m)\} \quad (50)$$

Se puede deducir que los datos están respectivamente etiquetados, donde:

- A la clase 1 le corresponde +1
- A la otra clase le corresponde el -1.

El TSVM ayudará a determinar 2 hiperplanos no paralelos:

$$H1 = x^T * w_1 + b_1 = 0 \quad (51)$$

$$H2 = x^T * w_2 + b_2 = 0 \quad (52)$$

En este caso, los datos de entrada provienen de una matriz A, es decir, que contiene a todas las muestras con cada una de sus variables. Por ejemplo:

$$A = [\text{x11 x12 x13};\ \text{x21 x22 x23}, \text{x31 x32 x33}]$$

Donde se interpreta: Existen 3 muestras (3 filas) y cada muestra tiene 3 variables (3 columnas) y cada muestra está representa un valor de salida 1 o -1 (y = 1, y = -1).

Para resolver el problema de TSVM, es necesario empezar analizando la forma primal del SVM para cada uno de los 2 planos. De acuerdo a ello, el problema de programación cuadrática para cada hiperplano presenta las siguientes ecuaciones:

TSVM1

$$Min_{(w_1, b_1, q_1)} \quad \frac{1}{2} \|Aw_1 + e_1 b_1\|_2 + C_1 e_2^T q_1 \quad (53)$$

Que está sujeto a:

$$-(Bw_1 + e_2 b_1) + q_1 \geq e_2$$

$$q_1 \geq 0$$

TSVM2

$$Min_{(w_2, b_2, q_2)} \quad \frac{1}{2} \|Bw_2 + e_2 b_2\|_2 + C_2 e_1^T q_2 \quad (54)$$

Que está sujeto a:

$$(Aw_2 + e_1 b_2) + q_2 \geq e_1$$

$$q_2 \geq 0$$

Donde:

- A -> Es la matriz con m1 ejemplos (filas) / muestras asociadas al hiperplano H1 (a la clase y=+1).
- B -> Es la matriz con m2 ejemplos (filas) / muestras asociadas al hiperplano H2 (a la clase de y=-1).
- W1 -> Es el vector de p1 parámetros del H1 de n columnas.
- W2 -> Es el vector de parámetros del H2 de n columnas.
- E1 -> Es el vector de "unos" (1 sola columna de "unos" de m1 filas).
- E2 -> Es el vector de "unos" (1 sola columna de "unos" de m2 filas).
- B1 -> Es el vector de parámetros del H1 (vector de coeficientes de intercepción).
- B2 -> Es el vector de parámetros del H2 (vector de coeficientes de intercepción).
- C1 -> Es el hiperparámetro, que compensa maximizar el margen o minimizar los errores asociado a la matriz A. C1>0.
- C2 -> Es el hiperparámetro, que compensa maximizar el margen o minimizar los errores asociado a la matriz B. C2>0.
- Q1 - > Es el vector de Errores de m1 filas.
- Q2 - > Es el vector de Errores de m2 filas.

Como se puede observar, el término b^2 ha sido adicionado a la función a minimizar, haciendo que esta función se vuelva estrictamente convexa (Fung, 2001) y facilitando la solución algebraica al momento de simplificar.

El primer término de la función a minimizar de TSVM1 es la suma de distancias al cuadrado que existe entre el hiperplano H1 y las muestras que pertenecen a la clase y=+1, esto quiere decir que, el hiperplano H1 tiende a estar cerca de las muestras que pertenecen a la clase y=+1. En tanto, las restricciones hacen que el hiperplano H1 se ubique a una distancia de por lo menos 1 de los puntos de clase y=-1.

El vector q1 se utiliza para medir el error. El segundo término de la función a minimizar de TSVM1, hace que se minimice la suma de las variables de error, es decir, trata de minimizar la clasificación errónea debido a los puntos que pertenecen a la clase y=-1. Además, el hiperparámetro C1 > 0 compensa la minimización de los 2 términos mencionados anteriormente.

Para llegar a la formulación dual de TSVM, se sigue el mismo proceso de la forma dual de SVM, pero en este caso para los 2 hiperplanos H1 y H2. Para ello, se toma como ejemplo la formulación dual de TSVM1.

Al utilizar la función auxiliar Lagrangiano en el problema de programación cuadrática de TSVM1, se obtiene la siguiente ecuación, donde existen 2 multiplicadores de LaGrange **α y beta.** :

$$L\,(w_1, b_1, q_1, \alpha, \beta) = \frac{1}{2}\,(A\,w_1 + e_1 b_1\,)^T\,(A\,w_1 + e_1 b_1) + \,C_1 e_2^T q_1 - \,\alpha^T(-(B\,w_1 + e_2 b_1) + \,q_1 - e_2) - \,\beta^T q_1 \quad (54)$$

Como el TSVM1 es un problema de optimización convexa, se pueden aplicar las condiciones de KKT, suficientes y necesarias, que se expresan de la siguiente manera:

$$A^T(A\,w_1 + e_1 b_1)B^T\alpha = 0 \quad (55)$$

$$e_1^T(A\,w_1 + e_1 b_1)e_2^T\alpha = 0 \quad (56)$$

$$C_1 e_2 - \,\alpha - \beta = 0 \quad (57)$$

$$-\,(Bw_1 + e_2 b_1) + q_1 \geq e_2 \quad (58)$$

$$\alpha^T\,(-(Bw_1 + e_2 b_1) + q_1 - \,e_2) = 0 \quad (59)$$

$$\beta^T q_1 = 0 \quad (60)$$

Donde: α ≥ 0, β ≥ 0, $q_1 \geq 0$

De la condición de la ecuación 55, se obtiene que: $C_1 e_2 - \,\alpha - \beta = 0$

Con el multiplicador de LaGrange Beta >=0 y A>=0, la condición mencionada resulta en α <= C1, por lo tanto, se obtiene:

$$0 \prec \alpha_i \prec C \quad (61)$$

Asimismo, de las condiciones (ecuaciones 55 y 56 respectivamente), se obtiene:

$$A^T(A\,w_1 + e_1 b_1)B^T\alpha = 0 \quad (62)$$

$$e_1^T(A\,w_1 + e_1 b_1)e_2^T\alpha = 0 \quad (63)$$

Al sumarlas y representarlas en forma de multiplicación de matrices, se obtiene:

$$[A^T \quad e_1^T]\;[A \quad e_1]\;[w_1 \quad b_1]^T \;+\; [B^T \quad e_2^T]\alpha = 0 \quad (64)$$

Por tanto, es posible redefinirlas como:

- $H = [A \quad e_1]$
- $G = [B \quad e_2]$
- $u = [w_1 \quad b_1]^T$

De esta manera, las condiciones quedarían expresadas con la siguiente ecuación:

$$(H^T H)u + G^T \alpha = 0 \tag{65}$$

En el caso de que H sea una matriz invertible, se expresaría como:

$$u = -(H^T H)^{-1} G^T \alpha \tag{66}$$

En el caso de que H no sea una matriz invertible, se introduce un término de regulación "e", tal como se indica en la investigación de Saunders (1998), multiplicando a la matriz de identidad I de dimensión apropiada. Este término regulador e*I, e>0 hará que la matriz ($H^T H$ + e*I) sea invertible, tal como se detalla a continuación:

$$u = -(H^T H + \epsilon I)^{-1} G^T \alpha \tag{67}$$

Esta ecuación solo podrá ser utilizada cuando sea necesario, mientras tanto, la demostración seguirá asumiendo que H es una matriz invertible. Por tanto, la forma dual de TSVM será expresada como:

$$Max\ L(w_1, b_1, q_1, \alpha, \beta) \tag{68}$$

De donde:

$$\nabla_{w_1} L\ (w_1, b_1, q_1, \alpha, \beta) = 0 \tag{69}$$

Derivando:

$$\frac{dL}{db_1} = 0 \qquad \frac{dL}{dq_1} = 0$$

$$\alpha \geq 0, \qquad \beta \geq 0$$

Ahora, usando las condiciones de KKT y la ecuación 67, se puede obtener:

DTWSVM1:

$$Max_\alpha \quad e_2^T \alpha - \frac{1}{2}\, \alpha^T G (H^T H)^{-1} G^T \alpha \tag{70}$$

Que está sujeto a:

$$0 \leq \alpha \leq C_1$$

Esta formulación dual del TSVM para el hiperplano 1, también es llamada "The Wolfe Dual" TSVM1. Para el hiperplano H2, se hará el mismo procedimiento implementado en el hiperplano H1, donde:

$$J = [A \quad e_1], \qquad Q = [B \quad e_2]$$

$$v = [w_2 \quad b_2]^T$$

$$v = (Q^T Q)^{-1} J^T v \tag{71}$$

Por tanto, "The Wolfe Dual" para el hiperplano H2 resulta de la siguiente manera:

DTWSVM2:

$$Max_v \quad e_1^T v - \frac{1}{2} v^T j (Q^T Q)^{-1} J^T v \tag{72}$$

Que está sujeto a:

$$0 \leq v \leq C_2$$

Finalmente, se resuelve "The Wolfe Dual" para cada hiperplano H1 yH2 y de esa forma, se obtienen los vectores **v** y **u;** donde:

$$u = [w_1, b_1]^T \tag{73}$$

$$v = [w_2, b_2]^T \tag{74}$$

Los vectores u y v determinarán los siguientes hiperplanos de separación:

- $x^T w_1 + b_1 = 0$
- $x^T w_2 + b_2 = 0$

Para determinar a qué clase pertenece cada nueva muestra (X), se debe resolver lo siguiente:

$$class\ (x) = arg_{i=1,2} Min\big(d_r(x)\big) \tag{75}$$

Finalmente, al despejar, se obtendría la siguiente ecuación:

$$d_r(x) = \frac{|x^T w_r + w_r|}{\|w^{(r)}\|} \tag{76}$$

Capítulo 3
Obtención y manejo de data

Para obtener las variables necesarias para esta investigación, se obtuvo data de entrada (variables medidas) a partir del 13 de noviembre del 2019 y data de salida (cantidad de trip – evaluación de 2 veces por semana para el cuidado de la plantación). Por ello, se plantearon algoritmos de predicción (de distintas técnicas), los cuales fueron mejorando con la cantidad de datos de entrada y de salida que se obtenía con el tiempo, hasta finalmente se utilizó la técnica de TSVM.

La data requerida fue tomada de la parcela que pertenece a la Cooperativa ASPROBO. Esta posee una forma cuadrangular de 1 hectárea y 466 metros de perímetro, cuya localización corresponde a las coordenadas 5° 16’ 13.4” S 79° 57’ 10.1” W, en el distrito de Buenos Aires, de la provincia de Morropón, tal como se aprecia en la figura 35.

Figura 35. Localización de la parcela a través de Google Earth
Fuente: Google Earth

La parcela, seleccionada como piloto para esta investigación, cuenta con una zona de empaque móvil ubicada a la entrada y en una posición que da a la zona de cultivo una forma de U (ver figura 37).

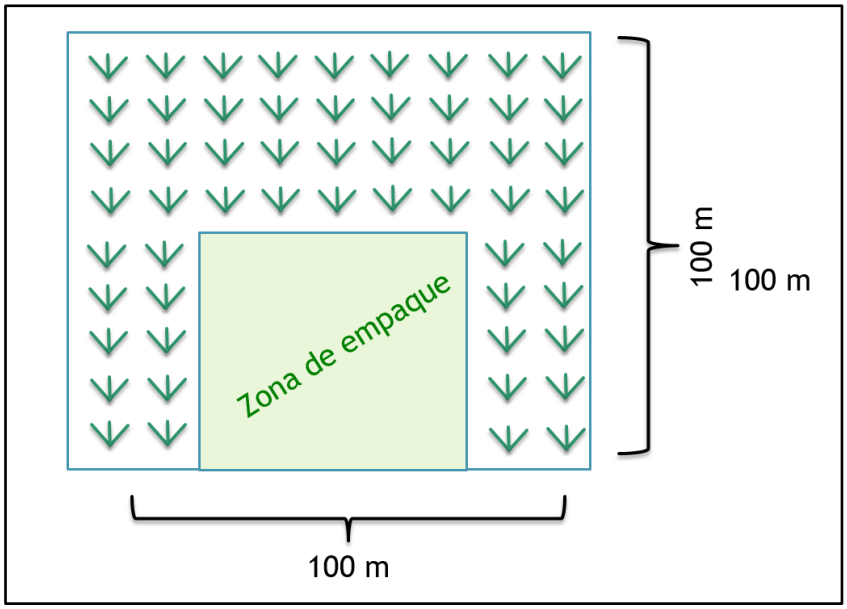

Figura 36. Representación gráfica de la parcela piloto

Esta data consistió en mediante equipos tecnológicos recopilar datos numéricos del clima, humedad, radiación solar, velocidad del viento, entre otras; la cual, se puede agrupar (o clasificar) en:

- datos del clima: proporcionados por la estación meteorológica ubicada por encima de las hojas del banano (aproximadamente a 4 metros de altitud).
- datos del microclima: proporcionados por los nodos (sensores ubicados entre las hojas y el suelo).
- datos del suelo: proporcionados por los nodos ubicados en el suelo, los cuales tuvieron tres medidas (30, 60 y 90 cm de profundidad).

3.1 Análisis de las plagas del banano

La evaluación de las plagas que afectan al banano es importante para controlar su propagación y para determinar que materiales utilizar para combatirlas; por tanto, la información de este proceso resulta elemental para la elaboración del algoritmo de predicción.

El método tradicional de la evaluación de plagas que se lleva a cabo en la Cooperativa ASPROBO, es realizado por el encargado de la evaluación de presencia de plagas en las parcelas y se ejecuta de la siguiente manera:

- En primer lugar, el encargado se dirige a cualquiera de los terrenos de la Cooperativa, bajo su propio criterio de observación, en el momento que estime conveniente.

- Al llegar a la parcela, realiza una selección de 25 plantas aleatorias para inspeccionarlas. La inspección es solo visual (algunas veces hace uso de un lente lupa “ojo de pescado”. Allí realiza un conteo del número de insectos encontrados o hallados a simple vista, a partir de los cuales, realiza cálculos para la respectiva evaluación de plagas.
- Después, la inspección realizada se apunta en una ficha, tal como se muestra en el anexo 1.
- Finalmente, con los datos de la ficha, se procede a realizar los cálculos respectivos, para completar la evaluación.

El formato o ficha contiene las principales plagas desarrolladas en la zona de ASPROBO, siendo la más importante para esta investigación, la plaga denominada “Trip de la mancha roja”. A continuación, se presenta la simulación de una evaluación del trip de la mancha roja, que ayudará a los productores para definir el estado de la plaga en la planta, la calidad del producto, el tiempo de maduración del fruto de la misma y, sobre todo, para especificar las medidas que se tomarán para el control de las mismas.

Los números de insectos encontrados en 25 plantas aleatorias de una misma parcela, se registran según el formato que se muestra en la figura 37.

FORMATO DE EVALUACIÓN DE PLAGAS DE BANANO

PRODUCTOR | EVALUADOR | FECHA

NOMBRE DEL PRODUCTOR																												
LADO	LD	LI	LD	LI	LI	LD	LD	LD	LD	LI	LI	LI	LD	LD	LI	LI	LD	LD	LD	LI	LI	LI	LI	LD	LD	LI	LI	LD

PLAGA	PLANTA / ESTRUC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	TOTAL	MAX	MIN	%
TRIPS DE LA MANCHA ROJA	NINFAS	2	3	1	4	2	1	3	2	1	5	2	2	3	2	1	4	5	2	3	6	5	4	2	1	3	69	6	1	2.8
	ADULTO	2	4	5	1	2	3	4	5	6	7	8	1	2	3	6	2	1	4	5	2	3	1	5	2	2	86	8	1	3.4
	TOTAL	4	7	6	5	4	4	7	7	7	12	10	3	5	5	7	6	6	6	8	8	8	5	7	3	5	155			6.2

Figura 37. Formato de evaluación de plagas

Después de realizar una suposición de los trips de la mancha roja encontrados para simular la evaluación de plagas, se ha determinado que la parcela inspeccionada cuenta con un porcentaje de 2.8% de ninfas y 3.4% de adultas de, lo que da un total de 6.2% de evolución de plaga en la parcela. Todos los porcentajes obtenidos han sido resultado de la suma de los números de insectos encontrados divididos entre las 25 plantas seleccionadas aleatoriamente.

Se debe aclarar que el trabajo realizado en la parcela es rápido y aproximado, es decir, al ser una sola persona la que debe revisar todas las parcelas asociadas, es probable que no se tome el tiempo necesario para examinar la planta en su totalidad; por tanto, lo que se suele hacer es aproximar la cantidad de insectos que ve por planta.

Para realizar los diagnósticos de condición por parcela, se efectúa el siguiente análisis, que considera un valor umbral de porcentaje de desarrollo.

- Para un % de 0.0% → No hay necesidad de fumigación.
- Para un % de 0.5% → Se aplica una fumigación preventiva (caldo sulfácico + aceite agrícola, crostitan y cantidades medidas de azufre).
- Para un % > 1.0% → Se aplica una fumigación curativa (Entrust SC.).

Para saber cada qué tiempo fumigar de manera orgánica y qué insumos utilizar, se debe realizar una evaluación previa de plagas, la cual consiste en llevar un conteo numérico de aquellos insectos visibles y, posteriormente, realizar un pequeño cálculo de suma y división; de esta forma, se obtiene un diagnóstico, que permite continuar con el proceso de control de plagas.

En la figura 38, al lado izquierdo se muestra un conteo rápido del encargado de inspeccionar la parcela y, al lado derecho, una “lupa pez de ojo”, que ayudan a observar los insectos encontrados en el tallo o en la hoja de la planta.

Figura 38. Conteo manual de plagas

Existen tres tipos de situaciones en las que el encargado puede encontrar plagas en una parcela:

- En 0.0%: significa que no se ha encontrado desarrollo de ninfas ni presencia de adultos, por lo tanto, no es necesario aplicar fumigación; sin embargo, debido a los eventuales ataques del trip de la mancha roja o de alguna otra plaga, se puede considerar el uso de caldo sulfácico como prevención.

- En un 0.5%: significa que existe presencia de crecimiento y desarrollo de la plaga y por ello, es necesario aplicar una fumigación preventiva, como el caldo sulfácico mezclado con aceite agrícola o el crostitan (mezcla de ajo, ají y azufre).

- En un % > 1.0%: significa que la plantación se encuentra en situación de peligro y/o riesgo, es decir, que la plaga no solo está en desarrollo y crecimiento, sino que, además está en reproducción y esto puede ser perjudicial para el fruto. Por tales motivos, se debe fumigar la planta con Entrust SC, un producto 100% orgánico con certificado OMRI que combina la baja toxicidad en los mamíferos de agente bilógicos y la alta eficacia de los insecticidas sintéticos.

Se debe tener en cuenta que, la aplicación de productos preventivos debe ser rotativa, es decir, se deben usar todos los productos y alternarlos entre las veces de fumigación, con el fin de que los insectos no se adapten al producto orgánico. En la figura 39, se puede observar un gráfico de aplicación de productos para el caso de la plaga en desarrollo.

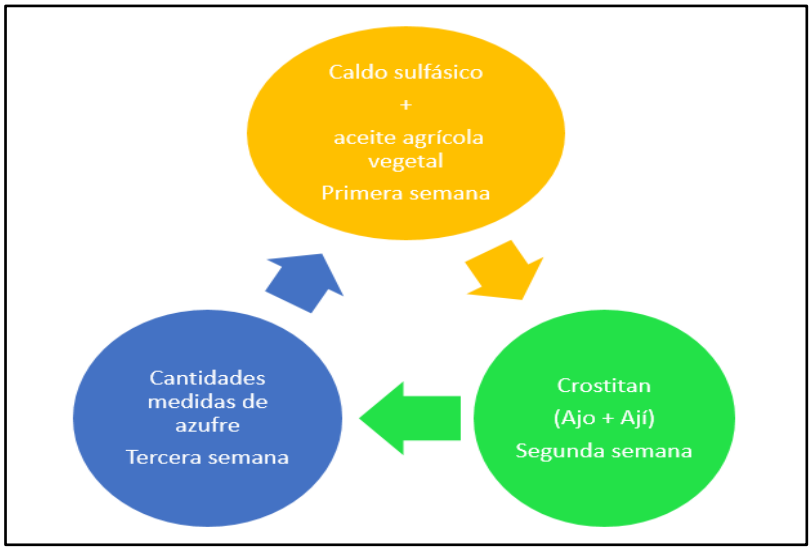

Figura 39. Aplicación de productos para control de plaga en la cuando el % = 0.5%

3.2 Instalación de equipamiento

Para obtener la data requerida, fue necesario instalar una estación meteorológica y sensores a lo largo de la parcela piloto, lo cual se llevó a cabo en los primeros días de noviembre del año 2019. Estos registraron tres niveles de variables, ya antes mencionados:

- Suelo (los nodos tienen 3 niveles de medida).
- Microclima (el ambiente que se forma desde el suelo hasta la parte inferior de las hojas del banano).
- Clima (encima de las hojas del banano).

3.2.1 *Estación meteorológica*

Al inicio del proyecto, se determinó que, debido a la forma rectangular de la parcela, el mejor criterio de ubicación es colocar la estación meteorológica al centro y los sensores, en los extremos, de forma que sea posible obtener una medida de ambas partes del terreno y parámetros que se ajusten a la realidad. Sin embargo, al momento de realizar la instalación de la estación meteorológica, se encontró que el punto sugerido estaba en medio de una planta y esto provocó que la estación se reubique alrededor del punto señalado en la figura 40.

Figura 40. Posición sugerida de la estación meteorológica

La estación meteorológica instalada (ver figura 41) constaba de un sistema *enviro Monitor*, que incluye un servidor (Gateway), que soporta dos nodos, los cuales están interconectados en topología malla, los cuales operan entre los 902 – 928 MHz, a través del protocolo de comunicación Zigbee. Cada uno de los nodos transmite la data recogida por los sensores al servidor, que se encarga de enviar la data a través de la conexión móvil en la red de datos, cuyo servicio es brindado por una operadora que tenga cobertura en la zona de instalación. En este caso el servicio de cobertura de red es brindado por Claro.

Figura 41. Estación meteorológica

3.2.2 *Sensores*

En la investigación previa se establecieron dos puntos que generarían la mejor toma de datos reales (ver figura 41); sin embargo, al ubicar los puntos en la parcela, resultó que uno estaba en el drenaje y el otro en medio de una plantación.

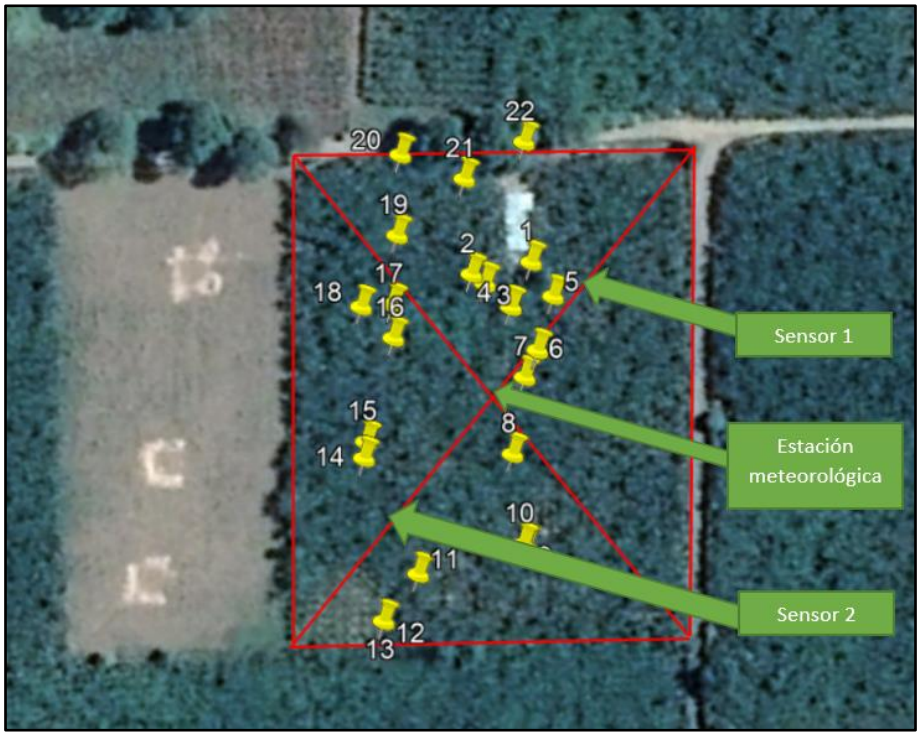

Figura 42. Ubicación ideal para los dos sensores y la estación meteorológica

Finalmente, se instalaron dos tipos de sensores:

- Sensor de temperatura y humedad del suelo GS3: constaba de 3 niveles (tal como se especificó líneas arriba, con una profundidad de 30, 60 y 90 cm). Tenía por finalidad, realizar la toma de lectura de la temperatura y humedad en el suelo. Su ubicación se detalla en la figura 42.

Figura 43. Sensores de temperatura y humedad para mediciones del suelo

- Sensor de temperatura y humedad del ambiente (microclima): estos sensores fueron los encargados de tomar las lecturas de temperatura y humedad relativa del ambiente que se encuentra por debajo de las hojas del banano (microclima). Se muestran en la figura 43.

Figura 44. Sensores de temperatura y humedad

3.3 Variables seleccionadas

Los nodos y la estación meteorológica instalados en la parcela, ofrecieron 36 variables de medición, que se detallan en las tablas 4 y 5.

Tabla 3. Variables meteorológicas - Nodo

	Sensor Microclima	**Unidades**
1	Temperatura baja	°C
2	Temperatura	°C
3	Temperatura alta	°C
4	Humedad baja	%
5	Humedad	%
6	Humedad alta	%
7	Punto de rocío	°C
8	Bulbo húmedo	°C
9	Índice de calor	°C
	Sensor de suelo	**Unidades**
10	Temperatura de suelo	°C
11	Humedad de suelo	%
12	EC de suelo	dS/m

Tabla 4. Variables meteorológicas - Estación central

	Variable	Unidades
1	Presión barométrica	mb
2	Temperatura	°C
3	Temperatura alta	°C
4	Temperatura baja	°C
5	Humedad	%
6	Punto de rocío	°C
7	Bulbo húmedo	°C
8	Velocidad del viento	m/s
9	Dirección del viento	--
10	Viento corriente	m
11	Alta velocidad del viento	m/s
12	Alta dirección del viento	--
13	Viento frío	°C
14	Índice de calor	°C
15	THW	°C
16	Lluvia	mm
17	Tasa de lluvia	mm/h
18	Evapotranspiración	mm
19	Grados día de calentamiento	--
20	Grados día de enfriamiento	--
21	THWS	°C
22	Radiación solar	W/m2
23	Energía solar	Ly

A pesar de tener diversas variables extraídas de los sensores colocados en los nodos y la estación meteorológica, se realizó una selección de variables de entrada, las cuales fueron elegidas según:

- El criterio de investigación del crecimiento, hábitat y ciclo de vida de los trips de la mancha roja.
- Investigaciones similares realizadas en años anteriores.
- Confesiones de los agricultores que conviven el día a día con el banano orgánico y la lucha para evitar el crecimiento de dicha plaga.

De la investigación del banano en Filipinas, realizada por el investigador Salvación, que está basada en la relación de rendimiento vs. clima, se obtuvo la hipótesis de que la producción del banano depende de la variación del clima; para ello, se utilizaron las variables que se presentan en la figura 45.

Table 1 – List of climatic variables used in yield modeling.

Variable	Abbreviation	Description	Unit
Annual Rainfall	AR	Total annual rainfall calculated from monthly CRU rainfall data	mm
Frequency of Wet Days	FWD	Total number of wet days (rainfall <1 mm) in a year	days
Precipitation Seasonality	PS	Coefficient of variation of monthly rainfall within a year	%
Annual Mean Temperature	AMT	Annual mean temperature	°C
Temperature Seasonality	TS	Amount of temperature variation within a year based on monthly temperature averages	°C
Annual Mean Diurnal Temperature Range	AMDR	The different between average annual maximum and minim temperature	°C

Note: CRU refers to the Climate Research Unit of the University of East Anglia (https://crudata.uea.ac.uk/cru/data/hrg/).

Figura 45. Variables utilizadas en la investigación de Salvacion
Fuente: Salvacion A., 2019

Otras investigaciones, también ayudaron en el proceso de selección de variables, priorizando a aquellas con mayor impacto en los algoritmos de predicción.

Finalmente, las variables seleccionadas fueron:

- Variable de Temperatura de microclima: obtenida de la estación meteorológica.
- Variable grado día: Esta variable de entrada se utilizó de la siguiente manera: Variable grado día = Temperatura de microclima – 11°; siendo 11° la temperatura mínima que resiste el trip de la mancha roja. Cabe indicar que el banano necesita de un clima caluroso, por lo que era poco probable encontrar temperaturas menores a 11°; sin embargo, se esto se determinó así, con la finalidad de garantizar la precisión del algoritmo.
- Variable velocidad del viento: variable de entrada obtenida de la estación meteorológica.
- Variable tasa de lluvia: Variable de entrada obtenida de la estación meteorológica.
- Variable Fumigación 1: Esta variable de entrada se obtuvo de los datos proporcionados por los dueños de la parcela, que hace referencia a la rutina de fumigación trimestral, que se realiza con productos orgánicos importados.
- Variable Fumigación 2: Esta variable de entrada se obtuvo de los datos proporcionados por los dueños de la parcela, que hace referencia a la rutina de fumigación diaria o interdiaria, que se realiza con productos como el calcio, el sodio, mezclas que incluyen al limón, entre otros.
- Variable limpieza: Esta variable de entrada hace referencia a la cantidad de días que transcurren desde que se hace la limpieza general de la parcela, para evitar la propagación del trip. Por lo general, esta limpieza está enfocada a la recolección de las hojas caídas o hijuelos que fueron eliminados.

- Variable logaritmo de la cantidad de Trips: Es la variable de salida representa la cantidad total de trips encontrados en la parcela, dividido entre la cantidad de plantas inspeccionadas.

Capítulo 4
Simulación y resultados

La investigación ha tenido diferentes cambios a lo largo de dos años, debido a la intención de encontrar la mejor solución al problema. Tal como se ha explicado en capítulos anteriores, la finalidad de esta investigación recae en encontrar un algoritmo de aprendizaje supervisado que se aproxime a la relación entre el crecimiento y reproducción de plagas con las condiciones climáticas.

Para ello, se han realizado cuatro algoritmos: dos de TSVM y dos de SVM, de los cuales, los algoritmos de TSVM arrojaron mejores resultados.

Para SVM1 y TSVM1, se utilizaron las siguientes variables:

Para la entrada:

- Variable grado día (Temperatura de microclima – 11°)
- Variables velocidad del viento
- Variable tasa de lluvia
- Variable fumigación 1 (acumulativa)
- Variable fumigación 2 (entre fumigaciones)
- Variable limpieza

Para la salida:

- Variable logaritmo de la cantidad de Trips

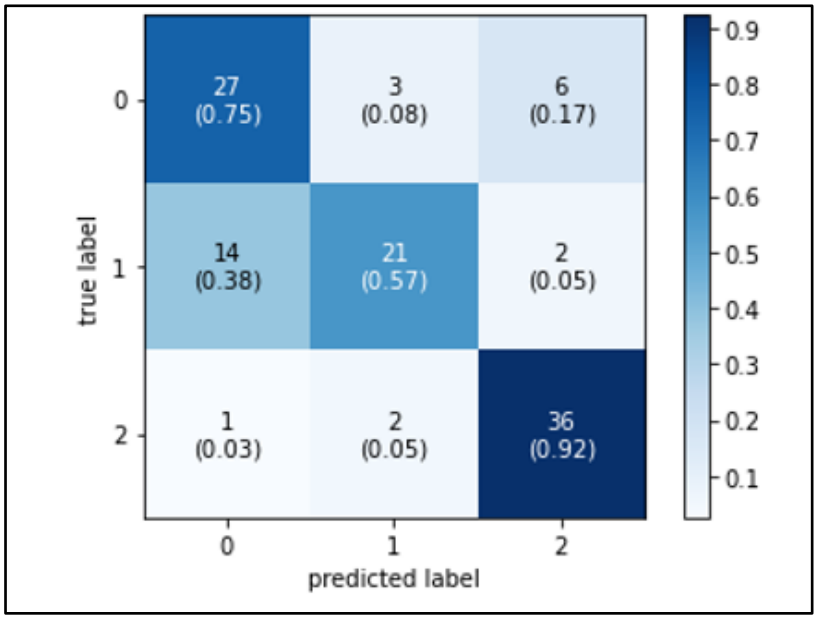

Figura 46. Matriz de confusión. Resultado SVM1 (grado día)

Para entender esta matriz de confusión, los valores de salida reales (0, 1 y 2) se encuentran en las coordenadas, y en las abscisas se encuentra las siguientes predicciones realizada por el algoritmo SVM1:

- Predicción 0 = 75%
- Predicción 1 = 57%
- Predicción 2 = 92%

La predicción total se podría considerar como la suma de todas las predicciones dividida entre 3, que resulta en 74.6%

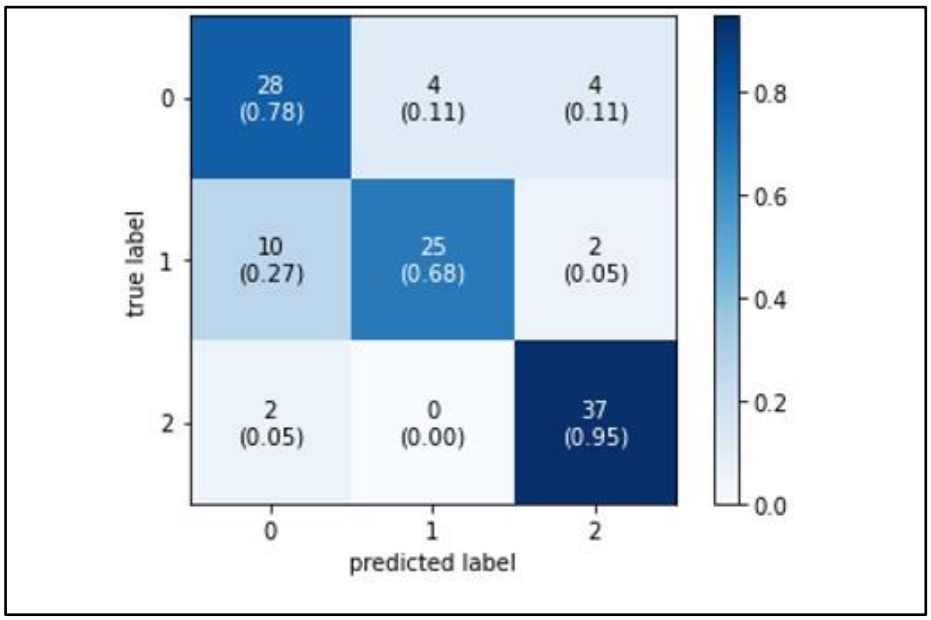

Figura 47. Matriz de confusión. Resultado TSVM1 (grado día)

En esta matriz de confusión, los valores de salida reales (0, 1 y 2) se encuentran en las coordenadas y en las abscisas, se encuentra la siguiente predicción realizada por el algoritmo TSVM1:

- Predicción 0 = 78%
- Predicción 1 = 68%
- Predicción 2 = 95%

La predicción total se podría considerar como la suma de todas las predicciones dividida entre 3, que resulta en 80.3%

Para SVM2 y TSVM2, se utilizaron las siguientes variables:

Para la entrada

- Variable de temperatura de microclima
- Variables velocidad del viento
- Variable tasa de lluvia
- Variable fumigación 1 (acumulativa)
- Variable fumigación 2 (entre fumigaciones)
- Variable limpieza

Para la salida:

- Variable logaritmo de la cantidad de Trips

Las salidas de los algoritmos están clasificadas en 3 (0, 1 y 2), siendo:

- 0 = poco conteo de trips en la parcela
- 1 = conteo moderado de trips en la parcela
- 2 = conteo elevado de trips en la parcela

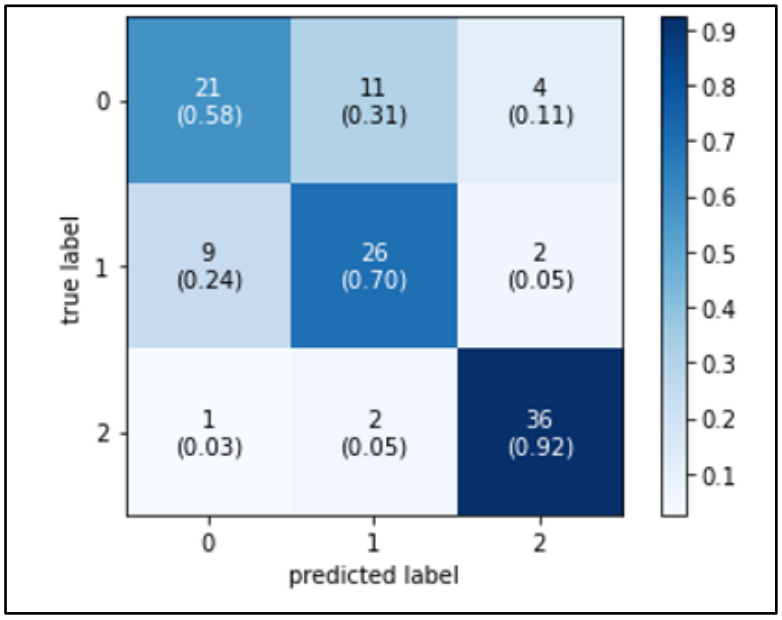

Figura 48. Matriz de confusión. Resultado SVM2 (temperatura de microclima)

En la matriz de confusión de la figura 49, los valores de salida reales (0, 1 y 2) se encuentran en las coordenadas se encuentras y en las abscisas, se encuentra la siguiente predicción realizada por el algoritmo SVM2:

- Predicción 0 = 58%
- Predicción 1 = 70%
- Predicción 2 = 92%

La predicción total se podría considerar como la suma de todas las predicciones dividida entre 3, que resulta en 73.3%

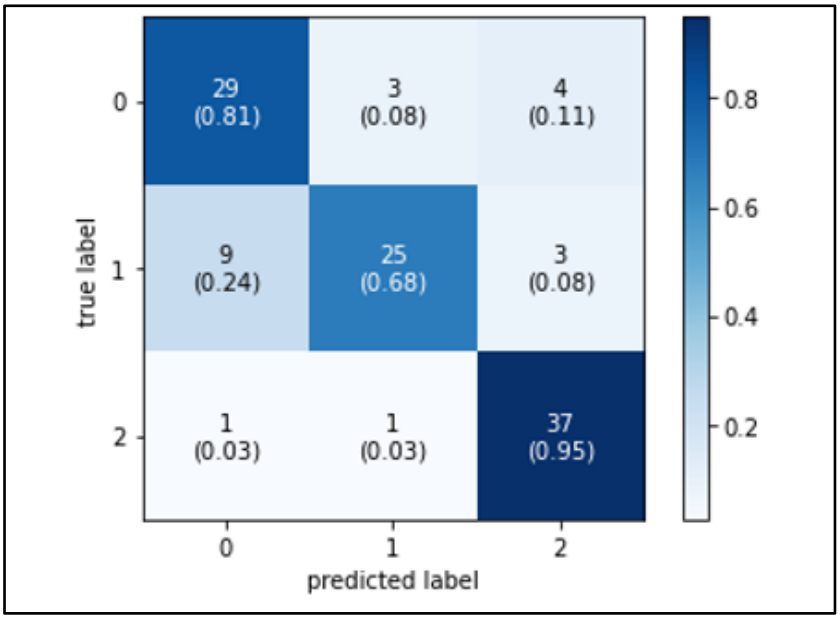

Figura 49. Matriz de confusión. Resultado TSVM 2 (temperatura de microclima)

En la matriz de confusión de la figura 50, los valores de salida reales (0, 1 y 2) se encuentran en las coordenadas se encuentras y en las abscisas, se encuentra la siguiente predicción realizada por el algoritmo TSVM2:

- Predicción 0 = 81%
- Predicción 1 = 68%
- Predicción 2 = 95%

La predicción total se podría considerar como la suma de todas las predicciones dividida entre 3, que resulta en: 81.3%

Los datos expresados anteriormente, se pueden agrupar en la tabla 6:

Tabla 5. Resumen de resultados en los códigos de predicción

ALGORITMO USADO	SVM1	TWINSVM1	SVM2	TWINSVM2
VARIABLE CAMBIANTE	GRADO DIA	GRADO DIA	TEMPERATURA DEL MICROCLIMA	TEMPERATURA DEL MICROCLIMA
PREDICCIÓN	**74.60%**	**80.30%**	**73.30%**	**81.30%**

UNIVERSITAS STUDIORUM PIURENSIS

Conclusiones

Aunque el estado del arte para esta tesis ofrecía un abanico de posibilidades (diferentes variables a seleccionar), en la práctica, ha resultado conveniente trabajar solo con las variables seleccionadas en el capítulo 4, pues estas facilitaban la ejecución del modelo de una forma más simple. Además, la cantidad limitada de data que estaba disponible, impidió aplicar algoritmos de predicción más avanzados (como las redes neuronales); por lo que se recurrió a los algoritmos de SVM y TSVM.

A inicios de la investigación, se presentó la hipótesis de que la variable más importante o con más impacto en la cantidad de trips era la variable grado día; sin embargo, con la variable de temperatura de microclima, a través del algoritmo TSVM2, se obtuvieron mejores resultados, demostrando que esta variable tiene mayor impacto en la proliferación de los trips de la mancha roja.

Adicionalmente, fue necesario fue necesario considerar las variables proporcionadas por los sensores instalados en la parcela y añadir las variables de limpieza y fumigación, ya que estas impactan directamente sobre los resultados.

Aparte de las variables que han sido tomado de los sensores que están instalados en la parcela ASPROBO, se vio necesario incluir la variable de limpieza y fumigación, ya que estas dos variables impactan directamente en la cantidad del trips de la mancha roja. Consecuentemente a esto, se ha tomado los días entre fumigaciones y entre limpiezas que se realizaron en la parcela para las variables de limpieza y fumigación.

El algoritmo que aportó las mayores predicciones fue el algoritmo *Twin Support Vector Machine* (TSVM), tanto al usar la variable temperatura de microclima, como la variable grado día, aportando mejores resultados (81.3% y 80.3% respectivamente).

Esta tesis busca sentar las bases para futuras investigaciones en el sector agrícola y especialmente, en el cultivo de banano orgánico, facilitando la predicción de la cantidad de trips de la mancha roja en base a variables climatológicas de fácil acceso.

El banano orgánico ha tenido popularidad durante las últimas décadas, por ello, los estudios realizados son pocos o nulos; esta investigación hasta su fecha de término no ha encontrado papers o análisis de técnica de clasificación con aprendizaje supervisado.

UNIVERSITAS STUDIORUM PIURENSIS

Referencias bibliográficas

Agricultura, O. d. (2017). Producción de banano orgánico en Perú. *FORO MUNDIAL BANANERO*.

Arys Carrasquilla - Batista, Alfonso Chacón - Rodriguez, Kattia Núñez - Montero, Olman Gómez - Espinoza, Johny Valverde, Maritza Guerrero - Barrantes. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el creciimiento microalgal. *Tecnología en Marcha, Encuentro de Investigación y Extensión 2016*.

Carrasquilla A, C. A. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. *Encuentro de Investigación y Extensión*, 33-45.

Castañeda, D. S. (2019). *APLICACIPON DE SUPPORT VECTOR MACHINE AL MERCADO COLOMBIANO.* Argentina.

Cortes, C. y. (1995). Support - Vector Networks. Machine Learning. 273-297.

Fung, G. M. (2001). *Proximal support vector machine classifiers.* In F. Provost & R. Srikant.

Ing. Juan Carlos Rojas Llanque. (2012). GUÍA TÉCNICA "ASISTENCIA TÉCNICA DIRIGIDA EN MANEJO INTEGRADO DE BANANO ORGÁNICO". *Agrobanco*.

Instituto de Investigación y Desarrollo de Comercio Exterior de la Cámara de Comercio de Lima - IDEXCAM. (2017). Bananas. *CÁMARA DE COMERCIO LIMA*.

Ipanaqué, W., Belupú, I., Estrada, C., Neyra, J., & Campos, J. (2021). *Producto 6. Monografía de diseño técnico de la apliacción.* Fontagro.

Jayadeva, K. R. (2007). Twin support vector machine for pattern clasification. *IEEE Trans Pattern Anal Mach Intell*, 905 - 910.

Juan Carlos Rojas Llanque. (2013). *"MANEJO INTEGRADO DE BANANO ORGÁNICO".* Pacanzga - Chepén - La Libertad.

Minagri: Exportación de banano orgánico peruano creció 94% en últimos 5 años. (2015). *Ministerio de Desarrollo Agrario y Riego*.

PIP Banano Orgánico. (s.f.). TRIPS DE LA MANCHA ROJA. *GOBIERNO REGIONAL PIURA GERENCIA REGIONAL DE DESARROLLO ECONÓMICO DIRECCIÓN REGIONAL DE AGRICULTURA*.

Ray. (s.f.). Effect of climate on provincial - level banan yield in the Philippines.

Reinoso, A. F. (2008). Asociaciones de pequeños productores y exportaciones de banano orgánico en el valle del Chira Piura. *Economía y Sociedad 69*.

Rosales, S. (10 de agosto de 2019). Perú tiene 170000 ha de plátano y banano orgánico en riesgo por plaga Fusarium. *GESTIÓN*.

Saunders, C. G. (1998). *Ridge regression learning algorithm in dual variables.*

Su, X. Y. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*.

Vapnik V. .N y Lerner, A. (1963). Pattern recognition using generalized portrait method. Automation and Remote Control. 774-78.

Zhu, X. &. (2009). Introduction to Semi-Supervised Learning.

Apéndices

UNIVERSITAS STUDIORUM PIURENSIS

Apéndice A. Código SVM temperatura

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler

# Clasificadores
from sklearn.svm import SVC

# Matrix de confusion y metricas
from sklearn.metrics import  confusion_matrix
from mlxtend.plotting import plot_confusion_matrix
from sklearn.metrics import classification_report

base_dataentrada = pd.read_csv("dataEntrada2.csv", sep=";")
base_datasalida = pd.read_csv("dataSalida2.csv", sep=";")

t_y = np.array(base_datasalida['t_y'])
y = np.array(base_datasalida['y'])

base_dataentrada['GradoDia'] = base_dataentrada['TempMic'] - 11

tempMic = np.array(base_dataentrada['TempMic'])
gradoDia = np.array(base_dataentrada['GradoDia'])
velViento = np.array(base_dataentrada['Vel Viento'])
tasaLluvia = np.array(base_dataentrada['TasaLluvia'])
t_F1 = np.array(base_dataentrada['tF1'])
t_F2 = np.array(base_dataentrada['tF2'])
t_L = np.array(base_dataentrada['tL'])

inputTempMic = []
inputGradoDia = []
inputVelViento = []
inputTasaLluvia = []
inputF1 = []
inputF2 = []
inputL = []
inputDeltaTiempo = []

for i in range(len(t_y)-1):
  k = np.linspace(t_y[i], t_y[i+1]-1, int(t_y[i+1]-t_y[i]))
  promTempMic = 0
  promVelViento = 0
  promTasaLluvia = 0
```

```
  acumGradoDia = 0
  degF1 = 0.044298
  degF2 = 0.59392462
  degL = 0.24856982

  for m in k:
    promTempMic += tempMic[int(m)]
    promVelViento += velViento[int(m)]
    promTasaLluvia += tasaLluvia[int(m)]
    acumGradoDia += gradoDia[int(m)]

  inputTempMic.append(promTempMic)
  inputGradoDia.append(acumGradoDia*(t_y[i+1]-t_y[i]))
  inputDeltaTiempo.append(t_y[i+1]-t_y[i])
  inputVelViento.append(promVelViento)
  inputTasaLluvia.append(promTasaLluvia)

  if t_F1[int(t_y[i])] > t_F1[int(t_y[i+1])]:
    inputF1.append((1-np.exp(-degF1*t_F1[int(t_y[i+1])]))/degF1)
  else:
    inputF1.append((np.exp(-degF1*t_F1[int(t_y[i])])-np.exp(-
degF1*t_F1[int(t_y[i+1])]))/degF1)

  if t_F2[int(t_y[i])] > t_F2[int(t_y[i+1])]:
    inputF2.append((1-np.exp(-degF2*t_F2[int(t_y[i+1])]))/degF2)
  else:
    inputF2.append((np.exp(-degF2*t_F2[int(t_y[i])])-np.exp(-
degF2*t_F2[int(t_y[i+1])]))/degF2)

  if t_L[int(t_y[i])] > t_L[int(t_y[i+1])]:
    inputL.append((1-np.exp(-degL*t_L[int(t_y[i+1])]))/degL)
  else:
    inputL.append((np.exp(-degL*t_L[int(t_y[i])])-np.exp(-
degL*t_L[int(t_y[i+1])]))/degL)

inc_Ant = y[0:112]
inc_Act = y[1:113]

clase = []
for sal in inc_Act:
    if sal <= 0.6383:
        clase.append(0)
    elif sal > 0.6383 and sal <= 1.5567:
        clase.append(1)
    else:
        clase.append(2)

data = pd.DataFrame()
data['Clase'] = clase
```

```
data['X0'] = inputDeltaTiempo   #DeltaTiempo
data['X1'] = inputTempMic       #tempMic*DeltaTiempo
data['X2'] = inputVelViento     #velViento*DeltaTiempo
data['X3'] = inputTasaLluvia    #tasaLluvia*DeltaTiempo
data['X4'] = inputF1            #efectoF1
data['X5'] = inputF2            #efectoF2
data['X6'] = inputL             #efectoL
data['X7'] = np.log(inc_Ant)    #log_Inc_ant

X_train = data.iloc[:,1:].values     # Valores de las variables de entr
ada
y_true = data.iloc[:,0].values       # Clases reales

scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)

kernel = 'rbf'
c_range = {'C' : [float(2**i) for i in range(-6,5)]}
gamma_range = {'gamma' : [float(2**i) for i in range(-
8,3)]} if kernel == 'rbf' else {}

param_range = {**c_range, **gamma_range}

cv_folds = 20
clf_SVC = SVC(kernel = kernel)
result = GridSearchCV(clf_SVC, param_range, cv = cv_folds, verbose=1, r
eturn_train_score = True)

result.fit(X_train, y_true)

best_mean_score = result.cv_results_['mean_test_score'][result.best_ind
ex_]
best_std_score = result.cv_results_['std_test_score'][result.best_index
_]

print("Best score: %.2f+-
%.2f" % (best_mean_score * 100, best_std_score * 100))

y_pred = result.best_estimator_.predict(X_train)
cm = confusion_matrix(y_true, y_pred)
fig, ax = plot_confusion_matrix(conf_mat=cm,
                                colorbar=True,
                                show_absolute=True,
                                show_normed=True)
plt.title('Matriz de Confusion', fontsize=15)

print(classification_report(y_true, y_pred))
```

```
def histograma(serie, nombre, xtitle, xl, xr, yb, yt):
    plt.figure(figsize=(7,5))
    df = pd.DataFrame(np.array(serie))
    suma = float(df.astype(float).sum())
    cont = int(df.astype(float).count())
    recta = suma/cont
    ax = sns.histplot(serie)
    plt.axvline(recta, linestyle="dashed", label="mean", color="k")
    plt.legend(loc="best",prop={"size":12})
    plt.title(nombre, size = 14)
    plt.xlabel(xtitle, size = 12)
    plt.ylabel('Count', size = 12)
    plt.xlim(xl,xr)
    plt.ylim(yb,yt)

histograma(np.array(result.cv_results_['mean_train_score']),'SVM Temper
atura Train', 'Accuracy', 0, 1, 0, 40)
histograma(np.array(result.cv_results_['mean_test_score']),'SVM Tempera
tura Test', 'Accuracy', 0, 1, 0, 40)
```

Apéndice B. Código TWIN SVM Temperatura de microclima

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler

# Clasificadores
import ltsvm.twinsvm as tsvm

# Matrix de confusion y metricas
from sklearn.metrics import  confusion_matrix
from mlxtend.plotting import plot_confusion_matrix
from sklearn.metrics import classification_report

base_dataentrada = pd.read_csv("dataEntrada2.csv", sep=";")
base_datasalida = pd.read_csv("dataSalida2.csv", sep=";")

t_y = np.array(base_datasalida['t_y'])
y = np.array(base_datasalida['y'])

base_dataentrada['GradoDia'] = base_dataentrada['TempMic'] - 11

tempMic = np.array(base_dataentrada['TempMic'])
gradoDia = np.array(base_dataentrada['GradoDia'])
velViento = np.array(base_dataentrada['Vel Viento'])
tasaLluvia = np.array(base_dataentrada['TasaLluvia'])
t_F1 = np.array(base_dataentrada['tF1'])
t_F2 = np.array(base_dataentrada['tF2'])
t_L = np.array(base_dataentrada['tL'])

inputTempMic = []
inputGradoDia = []
inputVelViento = []
inputTasaLluvia = []
inputF1 = []
inputF2 = []
inputL = []
inputDeltaTiempo = []

for i in range(len(t_y)-1):
  k = np.linspace(t_y[i], t_y[i+1]-1, int(t_y[i+1]-t_y[i]))
  promTempMic = 0
  promVelViento = 0
  promTasaLluvia = 0
```

```
  acumGradoDia = 0
  degF1 = 0.044298
  degF2 = 0.59392462
  degL = 0.24856982

  for m in k:
    promTempMic += tempMic[int(m)]
    promVelViento += velViento[int(m)]
    promTasaLluvia += tasaLluvia[int(m)]
    acumGradoDia += gradoDia[int(m)]

  inputTempMic.append(promTempMic)
  inputGradoDia.append(acumGradoDia*(t_y[i+1]-t_y[i]))
  inputDeltaTiempo.append(t_y[i+1]-t_y[i])
  inputVelViento.append(promVelViento)
  inputTasaLluvia.append(promTasaLluvia)

  if t_F1[int(t_y[i])] > t_F1[int(t_y[i+1])]:
    inputF1.append((1-np.exp(-degF1*t_F1[int(t_y[i+1])]))/degF1)
  else:
    inputF1.append((np.exp(-degF1*t_F1[int(t_y[i])])-np.exp(-
degF1*t_F1[int(t_y[i+1])]))/degF1)

  if t_F2[int(t_y[i])] > t_F2[int(t_y[i+1])]:
    inputF2.append((1-np.exp(-degF2*t_F2[int(t_y[i+1])]))/degF2)
  else:
    inputF2.append((np.exp(-degF2*t_F2[int(t_y[i])])-np.exp(-
degF2*t_F2[int(t_y[i+1])]))/degF2)

  if t_L[int(t_y[i])] > t_L[int(t_y[i+1])]:
    inputL.append((1-np.exp(-degL*t_L[int(t_y[i+1])]))/degL)
  else:
    inputL.append((np.exp(-degL*t_L[int(t_y[i])])-np.exp(-
degL*t_L[int(t_y[i+1])]))/degL)

inc_Ant = y[0:112]
inc_Act = y[1:113]

clase = []
for sal in inc_Act:
    if sal <= 0.6383:
        clase.append(0)
    elif sal > 0.6383 and sal <= 1.5567:
        clase.append(1)
    else:
        clase.append(2)

data = pd.DataFrame()
data['Clase'] = clase
```

```
data['X0'] = inputDeltaTiempo   #DeltaTiempo
data['X1'] = inputTempMic       #tempMic*DeltaTiempo
data['X2'] = inputVelViento     #velViento*DeltaTiempo
data['X3'] = inputTasaLluvia    #tasaLluvia*DeltaTiempo
data['X4'] = inputF1            #efectoF1
data['X5'] = inputF2            #efectoF2
data['X6'] = inputL             #efectoL
data['X7'] = np.log(inc_Ant)    #log_Inc_ant

X_train = data.iloc[:,1:].values     # Valores de las variables de entr
ada
y_true = data.iloc[:,0].values       # Clases reales

scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)

kernel = 'RBF'

c_range = {'C1': [float(2**i) for i in range(-6, 5)],
           'C2': [float(2**i) for i in range(-6, 5)]}
gamma_range = {'gamma' : [float(2**i) for i in range(-
8,3)]} if kernel == 'RBF' else {}

param_range = {**c_range, **gamma_range}

cv_folds = 20
clf_TSVM = tsvm.OVO_TSVM(kernel=kernel)
result = GridSearchCV(clf_TSVM, param_range, cv = cv_folds, verbose=1,
return_train_score = True)

result.fit(X_train, y_true)

best_mean_score = result.cv_results_['mean_test_score'][result.best_ind
ex_]
best_std_score = result.cv_results_['std_test_score'][result.best_index
_]

print("Best score: %.2f+-
%.2f" % (best_mean_score * 100, best_std_score * 100))

y_pred = result.best_estimator_.predict(X_train)
cm = confusion_matrix(y_true, y_pred)
fig, ax = plot_confusion_matrix(conf_mat=cm,
                                colorbar=True,
                                show_absolute=True,
                                show_normed=True)
plt.title('Matriz de Confusion', fontsize=15)
```

```
print(classification_report(y_true, y_pred))

def histograma(serie, nombre, xtitle, xl, xr, yb, yt):
    plt.figure(figsize=(7,5))
    df = pd.DataFrame(np.array(serie))
    suma = float(df.astype(float).sum())
    cont = int(df.astype(float).count())
    recta = suma/cont
    ax = sns.histplot(serie)
    plt.axvline(recta, linestyle="dashed", label="mean", color="k")
    plt.legend(loc="best",prop={"size":12})
    plt.title(nombre, size = 14)
    plt.xlabel(xtitle, size = 12)
    plt.ylabel('Count', size = 12)
    plt.xlim(xl,xr)
    plt.ylim(yb,yt)

histograma(np.array(result.cv_results_['mean_train_score']),'TSVM Tempe
ratura Train', 'Accuracy', 0, 1, 0, 175)
histograma(np.array(result.cv_results_['mean_test_score']),'TSVM Temper
atura Test', 'Accuracy', 0, 1, 0, 175
```

Anexos

UNIVERSITAS STUDIORUM PIURENSIS

Anexo 1. Formato de Evaluación de plagas de banano ASPROBO

FORMATO DE EVALUACIÓN DE PLAGAS DE BANANO							
PRODUCTOR		EVALUADOR		FECHA			

NOMBRE DEL PRODUCTOR							
LADO							
PLAGA	PLANTA / ESTRUC	1	2	3	…	24	25
TRIPS DE LA MANCHA ROJA	NINFAS						
	ADULTO						
	TOTAL						
ESCAMA	HOJAS 4ta						
	HOJAS 5ta						
	PEDÚNCULO						
	RACIMO - Mano superior - media - inferior						
	PSEUDOTALLO						
	HIJUELO						
	SEMANA (CINTA)						
	TOTAL						
COCHINILLA	PEDÚNCULO						
	RACIMO - Mano superior - media - inferior						
	PSEUDOTALLO						
	HIJUELO						
	SEMANA (CINTA)						
	TOTAL						
PICUDO	N° INDIVIDUOS / PL						
	SEMANA (CINTA)						
	TOTAL						
ARAÑITA ROJA	HOJAS PM						
	HIJUELO						
	MANIAS						
	SEMANA (CINTA)						
	TOTAL						
CERAMIDIA	HOJAS PM						
	HIJUELO						
	SEMANA (CINTA)						
	TOTAL						
PULGON	RACIMO						
	HIJUELO						
	PEDÚNCULO						
MOSCA BLANCA	HOJAS PM						
	HIJUELO						
	PSEUDOTALLO PM						
	SEMANA (CINTA)						