La investigación reportada en esta tesis es parte de los programas de investigación del CICESE (Centro de Investigación Científica y de Educación Superior de Ensenada, B.C.).

La investigación fue financiada por el CONACYT (Consejo Nacional de Ciencia y Tecnología).

Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Maestría en Ciencias
en Ciencias de la Computación

Predicción de plagas para el manejo de patologías en cultivos de vid

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Maestría en Ciencias

Presenta:

Berenice Martínez Téllez

Ensenada, Baja California, México

2022

Tesis defendida por

Berenice Martínez Téllez

y aprobada por el siguiente Comité

Dr. José Antonio García Macías
Codirector de tesis

Dr. Edgar Leonel Chávez González
Codirector de tesis

Dra. Mónica Tentori Espinosa

Dra. Rufina Martínez Hernández

Dr. Pedro Gilberto López Mariscal
Coordinador del Posgrado en Ciencias de la Computación

Dr. Pedro Negrete Regagnon
Director de Estudios de Posgrado

Resumen de la tesis que presenta Berenice Martínez Téllez como requisito parcial para la obtención del grado de Maestría en Ciencias en Ciencias de la Computación.

Predicción de plagas para el manejo de patologías en cultivos de vid

Resumen aprobado por:

Dr. José Antonio García Macías
Codirector de tesis

Dr. Edgar Leonel Chávez González
Codirector de tesis

El garantizar la seguridad alimentaria, la nutrición y promover agricultura sostenible, son parte del segundo objetivo de desarrollo sostenible de la agenda 2030. Para esto, los sistemas agroalimentarios deben incrementar su resiliencia ante las crecientes perturbaciones de orígenes diversos. La gestión del riesgo para perturbaciones que se pueden prevenir, por ejemplo las plagas y enfermedades, es fundamental para prever alteraciones importantes en los sistemas y así evitar las costosas intervenciones de recuperación. Dado que las plagas y enfermedades son una de las principales causas de la disminución de la calidad y cantidad de productividad agrícola, existe un reto encaminado a temprana detección para alertar a los agricultores, tomar ciertas precauciones y prevenir posibles pérdidas. Para abordar este reto, se han aplicado técnicas de aprendizaje de máquina con resultados prometedores. El presente trabajo de tesis presenta el primer estudio sobre la predicción de la presencia y los niveles de infestación de la plaga del piojo harinoso de la vid (PHV) en el estado de Baja California, con especial énfasis en el Valle de Guadalupe. Para lograr el objetivo se fusionó y procesó información de vigilancia a nivel de campo de los cultivos, así como información agroclimática para lograr descifrar la dinámica cultivo-clima-plaga. Los modelos empleados para la predicción fueron k vecinos cercanos (kNN) y el modelo HSP, ambas técnicas de aprendizaje de máquina basado en ejemplos. Con ambos modelos se obtuvieron métricas por encima del $73\,\%$ de clasificaciones correctas. El modelo HSP presentó ventajas por ser libre de parámetros. Por medio de la metodología seguida se logró identificar los parámetros que más impactan y propician el desarrollo de PHV; el mes, la fenología, la radiación solar total y la temperatura del suelo fueron las características más importantes las cuales determinan, en gran medida, el incremento de la población del PHV en el Valle de Guadalupe. La radiación solar total demostró ser un indicador de gran importancia; al alcanzar valores mayores a 600 cal/cm^2 se observa un claro incremento en la población del PHV. Con base en esta información, se recomienda la instalación de sensores de radiación en lugares estratégicos que permita tener un indicador para aplicar la técnica de control más adecuada.

Palabras clave: Predicción de plagas, piojo harinoso de la vid, aprendizaje de máquina, aprendizaje basado en ejemplos, k-vecinos cercanos, grafo de semi-espacios proximales

Abstract of the thesis presented by Berenice Martínez Téllez as a partial requirement to obtain the Master of Science degree in Computer Science.

Prediction of pests for the management of pathologies in vineyards

Abstract approved by:

Dr. José Antonio García Macías
Thesis Co-Director

Dr. Edgar Leonel Chávez González
Thesis Co-Director

Achieving food security, improving nutrition, and promoting sustainable agriculture is part of the second sustainable development goal of the 2030 agenda. For this, agri-food systems must become more resilient to increasing shocks of diverse origins. Risk management for preventable shocks, such as pests and diseases, is essential to anticipate major system disruptions and avoid costly recovery interventions. Since pests and diseases are one of the leading causes of the decrease in the quality and quantity of agricultural productivity, there is a challenge aimed at early detection signals to alert farmers, take certain precautions and prevent possible losses. Machine learning techniques have been applied to address above challenges, with promising results.This thesis work presents the first study on predicting the presence and infestation levels of the vine mealybug (VMB) in Baja California, with particular emphasis on the Guadalupe Valley. We merged and processed field-level crop monitoring information with agroclimatic information to decode the crop-climate-pest dynamics in this work. The models used for prediction were k nearest neighbors (kNN) and the HSP, both instance-based machine learning techniques. With both models, we obtained above $73\,\%$ of correct classifications. The HSP had the advantage of being a parameter-free learning algorithm. Through the methodology followed, it was possible to identify the parameters that most impact and favor the development of VMB; month, phenology, total solar radiation, and soil temperature were the most important characteristics that determine, broadly, the increase of the VMB population in the Guadalupe Valley.Total solar radiation proved to be an indicator of paramount importance; when radiation values hit more than 600 cal/cm^2, we observed an apparent increase in the VMB population. Based on this information, we recommend installing radiation sensors in strategic locations as the primary signal to apply the most appropriate control technique.

Keywords: Pest prediction, vine mealybug, machine learning, instance-based learning, k-Nearest Neighbors, half-space proximal graph

Dedicatoria

A mi familia.

Todo lo que soy se lo debo a ustedes. Gracias por su apoyo y amor incondicional en cada etapa y en cada decisión tomada, los amo ♡.

A mis amigos quienes me acompañan, apoyan y escuchan, todo es recíproco.

Agradecimientos

Al Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California, por la oportunidad y los medios para realizar mis estudios de posgrado.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de maestría. No. de becario: 1064998

Al Dr. Edgar Chávez y a la Dra. Rufina por sus invaluables recomendaciones para el presente trabajo, por su orientación y la confianza depositada.

A mi comité de tesis por sus comentarios y sugerencias que enriquecieron y mejoraron este trabajo.

Agradezco al Comité Estatal de Sanidad Vegetal de Baja California (CESVBC), en particular al Ingeniero Silvino de Jesús Aguilar y a todos los involucrados de la campaña contra plagas de vid por proporcionar la información analizada en el presente trabajo.

A Mich por ser la mejor compañía en en todo sentido de esta etapa. También a Mariana y Abraham por su calidez y calidad humana.

Tabla de contenido

Página

Lista de figuras

Figura Página

Lista de tablas

Capítulo 1. Introducción

La agricultura es la principal fuente de alimentos, materias primas y combustible; hacerla sustentable, incrementar la producción, reducir las pérdidas por plagas y desastres naturales contribuye al crecimiento económico de una nación (Altieri, 2018). Debido al crecimiento de la población se tiene una alta demanda de alimentos, lo cual ejerce una presión directa al sector agrícola. Existen condiciones que ponen en riesgo la seguridad alimentaria; condiciones como cambio climático, disminución de polinizadores, mala gestión del agua, plagas y enfermedades, por mencionar algunos.

El cambio climático afecta a los sistemas agroalimentarios a través de perturbaciones de corta duración; tales como fenómenos meteorológicos extremos, así como condiciones estresantes de aparición lenta como el aumento de temperaturas, la desertificación, la salinización y la pérdida de la biodiversidad (FAO, 2021). Hacerle frente al cambio climático es una tarea inevitable de toda la humanidad, las recomendaciones principales radican en desarrollar estrategias de adaptación, ya que el cambio climático determina cuánto alimento se puede producir y en dónde. Por otro lado, el sector agrícola constituye uno de los principales consumidores de agua dulce disponible a escala mundial, el crecimiento de las plantas depende en extremo de la disponibilidad de agua. Teniendo en cuenta el rápido agotamiento de los acuíferos y la poca recarga, es necesaria una gestión más eficaz del agua para conservarla mejor y lograr una producción agrícola sostenible (Benos et al., 2021). Por último, las plagas y enfermedades también ponen en riesgo la seguridad alimentaria por medio de las enormes pérdidas que provocan y del daño que, directa o indirectamente, provocan a los humanos; una plaga es cualquier organismo, planta o animal que tiene un efecto negativo sobre el sistema de producción agrícola y una enfermedad es una condición anormal de alteración del metabolismo de una planta. La práctica más utilizada en el control de plagas y enfermedades es la distribución de plaguicidas sobre la zona de cultivo. Esta práctica, aunque eficaz, tiene un alto costo económico y un importante costo medioambiental. Los impactos ambientales pueden ser residuos en los productos de cultivo, efectos secundarios en la contaminación de las aguas subterráneas, impactos en la fauna local y en los ecosistemas (Liakos et al., 2018). Ante este y otros problemas es necesario tomar medidas preventivas y optimizar las interacciones entre las plantas, los animales, los seres humanos y el medio ambiente, teniendo en cuenta al mismo tiempo los aspectos sociales que deben abordarse para lograr un sistema alimentario justo y sostenible (FAO, 2018). Existen estudios donde los insumos agroquímicos son dirigidos en términos de tiempo y lugar por medio del uso de métodos computacionales, un ejemplo de este caso es el de Meisner et al. (2016) quienes emplearon datos históricos de cultivos de algodón y de la gestión de plagas en el Valle de San Joaquín, California, para indicar el momento óptimo de aplicación de plaguicidas.

El aprendizaje de máquina (Machine Learning ML, por sus siglas en inglés) y cómputo de alto rendimiento crean nuevas oportunidades para decifrar, cuantificar y comprender diversos procesos en entornos operativos agrícolas. (Liakos et al., 2018). El aprendizaje automático es un subcampo de las Ciencias de la Computación y rama de la Inteligencia Artificial, el cual permite a un sistema *aprender* directamente de los datos, empleando algoritmos que se basan en una colección de ejemplos de algún fenómeno. La información puede proceder de la naturaleza, ser elaborados por humanos o generados por otro algoritmo (Burkov, 2019). Dicho de otro modo, el aprendizaje automático es el proceso de resolver un problema mediante la recopilación de un conjunto de datos y la construcción de un modelo estadístico que permita predecir un aspecto particular. Típicamente el proceso de aprendizaje se lleva a cabo con un conjunto de datos, llamados datos de entrenamiento; cada uno de estos ejemplos o instancias está descrito mediante un conjunto de atributos, características o variables. Los atributos pueden ser nominales, binarios, ordinales o numéricos. Una vez finalizado el proceso de aprendizaje, el modelo puede utilizarse para clasificar, predecir o agrupar nuevos ejemplos utilizando lo aprendido durante el proceso de entrenamiento (ver fig. 1).

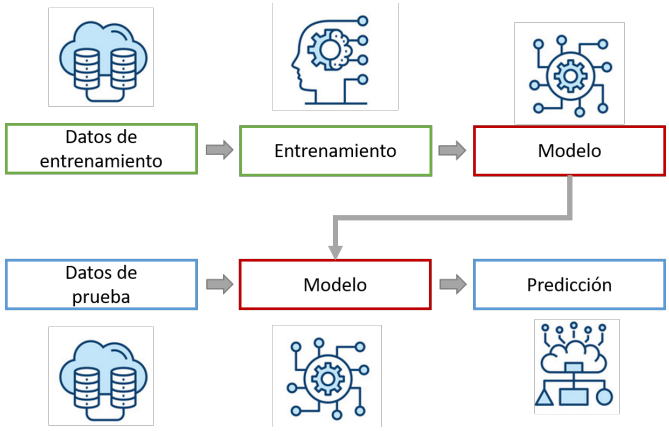

Figura 1. Enfoque clásico del aprendizaje automático.

La aplicación de técnicas de aprendizaje automático en agricultura brinda la oportunidad de realizar una agricultura más eficiente y precisa con menos mano de obra así como de una producción de alta calidad (Hasan et al., 2020). Por ejemplo, Han et al. (2020) analizaron datos climáticos, de teledetección y del suelo para predecir el rendimiento del trigo de 1 a 2 meses antes de las fechas de cosecha. Xenakis et al. (2020) presentan un sistema de apoyo al diagnóstico de enfermedades de plantas que utiliza una plataforma de Internet de las cosas para controlar un sistema robótico ligero; este sistema aplica un algoritmo de redes neuronales convolucionales para realizar el diagnóstico y la clasificación temprana de enfermedades de las plantas. Las ganancias de este tipo de estudios son obtenidas principalmente al proporcionar recomendaciones e información sobre los cultivos, beneficiando así a los agricultores para minimizar las pérdidas en la agricultura (Meshram et al., 2021).

Liakos et al. (2018) y Benos et al. (2021) identifican cuatro categorías de aplicaciones del aprendizaje de máquina en agricultura (Tabla 1); la gestión de los cultivos, del agua, del suelo y del ganado. La gestión de cultivos representa el mayor porcentaje de estudios y se subdivide en otras subcategorías: predicción del rendimiento, detección de enfermedades, detección de malas hierbas, reconocimiento de cultivos y calidad de los cultivos.

Tabla 1. Cuatro categorías genéricas en la agricultura que explotan las técnicas de aprendizaje automático con sus porcentajes de aportación de estudios realizados de acuerdo con Benos et al. (2021)

Categoría	Descripción	Sub categorías
Gestión del cultivo 61 %	Incluye aspectos versátiles que se originan en la combinación de técnicas agrícolas para gestionar el entorno biológico, químico y físico de los cultivos, con el fin de alcanzar objetivos cuantitativos y cualitativos.	a) Rendimiento del cultivo b) Detección de enfermedades c) Detección de malas hierbas d) Reconocimiento de cultivos e) Calidad de los cultivos
Gestión del agua 10 %	Ofrece el potencial del riego de tasa variable para lograr un ahorro de agua. Se puede lograr aplicando el riego con tasas que varían según las necesidades específicas de las distintas zonas de gestión.	-
Gestión del suelo 10 %	Una correcta gestión del suelo es vital para conocer el potencial de la tierra, para prevenir la degradación, frenar el desequilibrio suelo-nutrientes y para minimizar la erosión del suelo.	-
Gestión del ganado 19 %	Aplicar técnicas para controlar la salud de los animales en tiempo real y reconocer los mensajes de alerta. Mejorar la producción en fases iniciales. Controlar la calidad y las condiciones de vida de los animales.	a) Bienestar de los animales b) Producción ganadera

Aunque se han tenido avances considerables en cada categoría, existen varios problemas en proceso, los principales están relacionados con la instalación de sensores en los campos agrícolas por diversas razones, por ejemplo, altos costos de las tecnologías de información y comunicación, las prácticas tradicionales y la falta de información (Benos et al., 2021). Adicionalmente los conjuntos de datos disponibles no reflejan casos realistas, ya que son generados por pocas personas que obtienen imágenes o muestras en periodos cortos de tiempo y en áreas muy limitadas, por lo que es necesario tener más conjuntos de datos recolectados en campo. Además, debe fomentarse el desarrollo de técnicas de ML mediante la incorporación del conocimiento de los expertos de las diferentes áreas, por ejemplo, ciencias de la computación, agricultura y sector privado, con el objetivo de diseñar soluciones realistas (Benos et al., 2021).

Con relación a la detección de enfermedades en cultivos existe una amplia serie de publicaciones (Fe-

rentinos, 2018; Nanni et al., 2020; Shah et al., 2021, etc.) las cuales consisten en aplicar técnicas de aprendizaje profundo a un conjunto masivo de imágenes para identificar el cultivo, la enfermedad y en ocasiones la severidad. Por otro lado, con relación a la detección de plagas por medio de imágenes las publicaciones se reducen: Liu y Wang (2020) y Li et al. (2020) quienes emplearon imágenes RGB para reconocer enfermedades y plagas que afectan a cultivos de tomate y arroz, respectivamente. Si bien el uso de aprendizaje profundo es de gran utilidad, sigue sin resolverse el problema de brindar recomendaciones oportunas sobre la posible aparición de alguna plaga o enfermedad, además de que no siempre reflejan casos prácticos. Ante este escenario es que se opta por trabajar con técnicas de aprendizaje automático empleando información existente recolectada en campo por el personal adecuado. Dado que las condiciones climatológicas condicionan parcialmente la aparición, intensidad y crecimiento de una plaga y con información también de las condiciones agroclimáticas. Por ejemplo, temperatura ambiental y del suelo, humedad relativa, evotranspiración, precipitación, etc.

Por otra parte, el buen funcionamiento de los programas fitosanitarios es de vital importancia para prevenir la introducción de plagas exóticas, entre otras funciones. Las especies invasoras o plagas exóticas se refieren a insectos, microbios, enfermedades, plantas o malezas que no son oriundas de la región, sino que fueron introducidas en una determinada zona en la cual pudo reproducirse y desarrollarse mas allá de los límites normales; invaden y se establecen en áreas nuevas provocando graves daños ecológicos y económicos. Un gran ejemplo es la plaga filoxera, la cual se considera como la plaga más devastadora en la historia de la viticultura mundial. Se trata de un insecto de origen americano que fue introducido en Europa en el año 1863 por plantas importadas de Estados Unidos y acabó con más de cinco millones de hectáreas de viñedos en Europa entre 1870 y 1930 (Martínez, 2018). El insecto parasita la vid americana sin dañarla pero al ser transferido a la vid europea acaba con las raíces provocando su muerte. El control de la filoxera se basa en el injerto de variedades europeas sobre variedades resistentes, por ejemplo, vides americanas resistentes al insecto.

Con lo mencionado anteriormente, es posible resaltar que se requieren desarrollar soluciones que estén en armonía con el ambiente y con el ser humano, además que sean escalables, es decir, que puedan ser aplicadas en grandes extensiones usando los recursos proporcionadamente. Es imprescindible también tomar ventaja de la información que constantemente se esta generando en el sector agrícola, que sea obtenida y almacenada por diversos sensores o por monitoreos de personal en campo. Aquellas soluciones que empleen técnicas de aprendizaje de máquina, métodos basados en los datos, inferencia, automatización y reconocimiento de patrones, son el objeto del presente trabajo.

Problema a resolver

De manera sucinta el problema a resolver es *predecir con grado aceptable de confianza la incidencia de una plaga de interés en un cultivo específico*. Para alcanzar este objetivo ciertos insumos y pasos son necesarios:

1. Disponer de información de las condiciones agroclimáticas provenientes de estaciones meteorológicas (temperatura, humedad relativa, precipitación, etc). Así como también información procedente de muestreos en campos de cultivo para una plaga determinada (fenología, variedad, presencia de plaga, nivel de infestación, etc).
2. Los insumos anteriores se convierten en variables de un sistema, nuestro objetivo es detectar cuáles de ellas determinan el comportamiento del fenómeno que estamos observando, en este caso, la incidencia de una plaga.
3. Una vez identificadas las variables, tendremos una hipótesis sobre las condiciones que se tienen que cumplir para que aparezca la plaga.
4. Utilizar un modelo estadístico basado en datos que brinde información acerca de la evolución de la plaga y a una posible gestión.
5. La información anterior servirá para brindar recomendaciones a los tomadores de decisiones.

El caso de estudio que se aborda en el presente trabajo corresponde a una plaga introducida conocida como Piojo Harinoso de la Vid (PHV). La especie de mayor importancia encontrada en México y en Estados Unidos es la *Planococcus Ficus*, la cual es proveniente de Egipto e Israel. Se registró por primera vez en México en el estado de Sonora en el año 2000 y en Baja California en 2014. El atacar esta plaga es clave en las regiones vitivinícolas ya que representa hasta un $30\,\%$ de pérdidas de la producción total (Agricultura-Senasica, 2021), cambia las propiedades organolépticas de los vinos, es vector de virus asociado con diversas enfermedades (enrollamiento de la hoja, mancha roja), entre otros daños. En el estado de Baja California, en particular, en el Valle de Guadalupe, Ensenada, se tienen altas tasas de incidencia y consecuentemente, graves daños. Ante este escenario se han llevado a cabo campañas contra las plagas de la vid y se formó un grupo técnico del piojo harinoso de la vid, el cual realiza muestreos y trampeos con una metodología bien definida: el monitoreo se realiza en 3, 800 hectáreas, se instalan trampas con feromonas para capturar a los machos, las cuales se revisan cada 15 días; además se realiza un muestreo aleatorio constante en lotes de 10 hectáreas, donde se revisan 75 plantas aleatorias en las

que se cuenta el número de piojos y el número de plantas infestadas. Con esta información se determina la presencia y el nivel de infestación de cada lote, calificándolo en cuatro niveles de infestación: sin presencia, nivel leve, nivel medio y nivel fuerte. Adicionalmente el personal en campo anota información relevante del cultivo, por ejemplo, la variedad, la etapa fenológica, el tipo de daño encontrado, etc. Por otro lado, se extrae la información agroclimática diaria de las estaciones meteorológicas cercanas a cada lote, por medio de la ubicación y de la fecha es posible combinar la información que empleará para analizar la correlación que existe entre la plaga, cultivo y condiciones climáticas.

Con los conjuntos de datos recolectados y fusionados se aplicaron diversas técnicas de limpieza y exploración para analizar y transformar la información. Los datos procesados y transformados en un formato adecuado se ingresan a un modelo de aprendizaje que se encarga de reconocer los patrones entre las características y la variable objetivo (nivel de infestación). Previo al proceso de entrenamiento se aplicaron diversas técnicas de selección de características, para reducir las variables de entrada al modelo utilizando sólo las características relevantes y eliminando el ruido en los datos. Algunas de las variables que resultaron ser más relevantes fueron: mes, etapa fenológica, radiación y temperatura. Se probaron dos modelos de aprendizaje de máquina: k Vecinos Cercanos (kNN) y Half Space Proximal (HSP). kNN y HSP son algoritmos de aprendizaje basado en ejemplos, este tipo de aprendizaje supervisado emplea los ejemplos almacenados en la base de datos para compararlos directamente con las nuevas consultas. Los algoritmos son basados en instancias, y estrictamente no construyen un modelo, solo almacenan la información y el procesamiento se realiza al momento de cada nueva consulta. El conjunto de entrenamiento está etiquetado y cada etiqueta corresponde a un objeto. El algoritmo kNN consiste en calcular la distancia entre una consulta y cada uno de los ejemplos dentro del conjunto de entrenamiento, se seleccionan los k objetos más cercanos y la etiqueta de la consulta será la más votada. La calidad de la clasificación depende del parámetro k, el cual se determina usualmente de manera experimental, este valor es de vital importancia para obtener buenos resultados. Como ejemplo de lo dicho, se observa en la figura 2 cómo al cambiar ligeramente el valor de k la consulta realizada pertenece a una clase u otra. Diversas propuestas se han generado para escoger automáticamente el valor de k, en el estado del arte se encuentran los métodos *kTree* y *k*Tree*, donde se agrega una etapa previa al entrenamiento del algoritmo kNN para encontrar los valores óptimos de k para cada consulta (Zhang et al., 2017).

Ante la limitante mencionada sobre la elección del valor k, Talamantes y Chavez (2022) formularon un algoritmo también basado en ejemplos, pero sin la necesidad de elegir dicho valor. El algoritmo HSP toma la ventaja del grafo de semiespacios proximales y en lugar de comparar los objetos más cercanos a la consulta, se comparan los objetos obtenidos por el algoritmo del grafo HSP (ver Fig. 3), posteriormente

se aplica una regla de votación para realizar la clasificación.

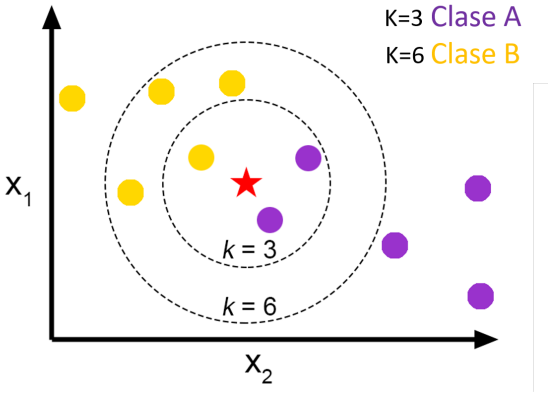

Figura 2. Funcionamiento y principal problema del algoritmo kNN.

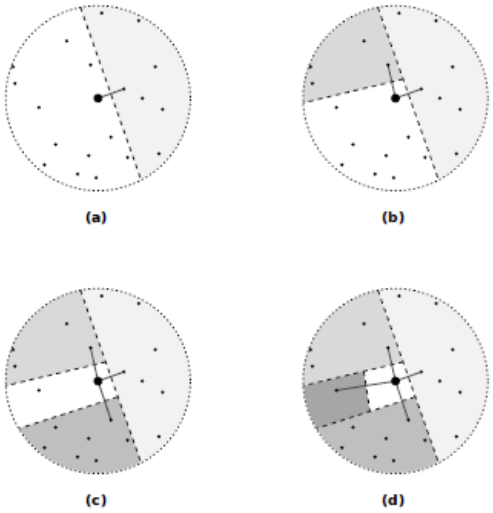

Figura 3. Selección de vecinos para una consulta (Talamantes y Chávez, 2021).

1.1. Antecedentes

La agricultura tradicional, sin tecnología de escala industrial, es practicada en gran parte del mundo. Aunque se tiene la existencia de tractores autónomos, se sigue circunscribiendo a labores mecánicas. La parte que nos interesa, es las tecnologías de la información aplicadas a la agricultura.

Existe un término conocido como agricultura 4.0, Santos Valle y Kienzle (2020) la definen como "Una agricultura que integra una serie de innovaciones para producir productos agrícolas. Estas innovaciones engloban la agricultura de precisión, el internet de las cosas (IoT) y macrodatos para lograr una mayor eficiencia en la producción", señalan que dentro de este paradigma la digitalización, automatización y la inteligencia artificial (IA) juegan un papel importante en la producción de cultivos incluyendo el control de malas hierbas y el control de plagas.

El control de plagas ha existido desde que inició la agricultura y ha tenido diferentes estrategias o enfoques. En la actualidad, la estrategia más utilizada es a través del uso de plaguicidas. Sin embargo, a pesar de la popularidad de las sustancias químicas, existen muchos riesgos ambientales y para la salud humana asociados a su uso (Jurado, 2020). En este sentido, se tiene la opción de realizar un Manejo Integrado de Plagas (MIP), estrategia que provee a los productores de herramientas y planes para vigilar y controlar plagas, reduciendo considerablemente el uso de plaguicidas, que generalmente son costosos y peligrosos, siendo el monitoreo constante la base de cualquier MIP.

Figura 4. Comparación entre agricultura convencional y agricultura 4.0 (Santos Valle y Kienzle, (2020)).

Dentro del MIP se puede mencionar el uso del control biológico que es una tecnología que aprovecha a los enemigos naturales de las plagas con la idea de reducir las poblaciones sin afectar las producciones agrícolas. Sin embargo, existen situaciones donde el control biológico no es una opción a seguir dada la diversidad de ecosistemas, cultura y economía, por lo que se crean programas específicos para conocer otras opciones para el control de plagas. Dentro de estos programas específicos se busca visibilizar la eficacia y conveniencia del uso de tecnologías que permitan lograr un manejo agroecológico en el futuro, en donde se reduzca al mínimo indispensable el uso de plaguicidas para la producción agrícola (Jurado, 2020).

La predicción de plagas y enfermedades empleando técnicas de aprendizaje de máquina no tiene un origen claro, existen estudios desde los 70's donde aplicaron pruebas estadísticas y regresiones múltiples para intentar predecir o pronosticar enfermedades en las plantas (Bourke, 1970; Coakley, 1988; Ishiguro y Hashimoto, 1991). Las primeras referencias bibliográficas que utilizan aprendizaje de máquina aparecen en los 90's, con un incremento en el año 2000. Algunos ejemplos son los trabajos de Chakraborty et al. (2004) y Kaundal et al. (2006), quienes explícitamente emplearon diversas técnicas de aprendizaje de máquina para predecir enfermedades en leguminosas y en arroz respectivamente.

Recientemente se tiene el estudio nacional realizado por Rodríguez-Moreno et al. (2020), en el cual empleando un modelo de árboles de clasificación y regresión (CART's por sus siglas en inglés) se analizaron los datos meteorológicos y datos de incidencia y severidad de la roya de trigo para predecir la presencia/ausencia de patógenos en cultivos del estado de Sonora, México.

1.2. Justificación

México cuenta con un territorio nacional de 196 millones de hectáreas de las cuales 145 millones (el $73\,\%$) se dedican a la actividad agropecuaria. Cerca de 30 millones de hectáreas son tierras de cultivo y 115 millones son de agostadero (Ramírez, 2016). El Instituto Nacional de Estadística y Geografía (INEGI) estima que en México hay 5.5 millones de personas dedicadas al trabajo agrícola, y de acuerdo con la Encuesta Nacional Agropecuaria del 2019 (ENA, 2019) solo el $37.7\,\%$ de las unidades de producción utiliza alguna TIC (Tecnología de la Información y Comunicación); una unidad de producción se refiere al conjunto de terrenos, infraestructura, maquinaria y equipo, animales y otros bienes utilizados en las actividades agropecuarias. De este porcentaje la tecnología más usada es el teléfono celular, utilizado por $88.1\,\%$ de las unidades de producción, seguido por el teléfono fijo con $19.8\,\%$. Mientras que el Internet sólo era aprovechado por el $7.9\,\%$ de las unidades de producción (ENA, 2019). En ocasiones esto dificulta actualizaciones e implementaciones tecnológicas en las diversas actividades agropecuarias del país, de hecho, los tres principales problemas presentados durante el desarrollo de las actividades agropecuarias reportados son: altos costos de insumos y servicios, dificultad para la comercialización, y falta de capacitación y asistencia técnica.

De acuerdo con la FAO hasta un $40\,\%$ de la producción agrícola mundial se pierde por causa de las plagas que llegan a afectar a los diferentes cultivos (FAO, 2021). En México el porcentaje es similar, teniendo entre el $20\,\%$ y $40\,\%$ de pérdidas de la producción de alimentos debido a las plagas o las enfermedades de los cultivos. En el territorio nacional se tiene la presencia de muchas plagas y enfermedades como son malezas, hongos, bacterias, virus, nematodos, insectos, aves, roedores, etc. De manera más específica y como ejemplos existe gran variedad de pulgones que afectan a cultivos de papa, tomate, chile y algodón. Algunos escarabajos conocidos como gallinas ciegas afectan a cultivos de maíz y sorgo principalmente. La mosca de la fruta provoca grandes disminuciones en la calidad de productos frutícolas. El virus rugoso del tomate y el psílido asiático de los cítricos son otros ejemplos que provocan graves daños y pérdidas en el territorio nacional (Jurado, 2020).

El estado de Baja California es el estado con mayor producción de uva industrial, y de acuerdo con la Secretaría de Agricultura y Desarrollo Rural, cuenta con el $87\,\%$ de empresas vitivinícolas del país. La plaga del piojo harinoso de la vid se registró en territorio nacional por primera vez en el año 2000 en estado de Sonora y en 2014 en Baja California. Ante los daños que provoca esta plaga, el Servicio Nacional de Sanidad, Inocuidad y Calidad Agroalimentaria (SENASICA) y la Secretaría del Campo y Seguridad Alimentaria (SCSA), han generado diversos programas; *Campaña contra plagas de la vid, Plan de trabajo para el manejo del piojo harinoso de la vid, Control biológico para el manejo de las*

plagas de la vid y la formación del *Grupo técnico del del Comité Estatal de Sanidad Vegetal de Baja California (CESVBC)*, entre otras iniciativas. Estos programas tienen objetivos diversos, por ejemplo; desarrollar estrategias operativas para muestreos y trampeos para una detección oportuna del PHV, realizar monitoreos constantes, brindar soluciones con los diferentes tipo de control existentes, realizar una caracterización fitosanitaria de la vid, generar estrategias para el control de hormigas, elaboración de manuales para el manejo fitosanitario del PHV, elaboración de materiales divulgativos para productores, entre otras actividades (CICESE, 2022).

El grupo técnico del Comité Estatal de Sanidad Vegetal de Baja California (CESVBC) lleva a cabo visitas constantes a más de 200 ranchos, abarcando un área de 3, 800 hectáreas, para realizar muestreos, trampeos y monitoreo desde el año 2017. El monitoreo se realiza a través de dos sistemas; la colocación de trampas con feromonas y el muestreo. Las trampas se colocan en áreas sin presencia de PHV y se revisan cada 15 días; el muestreo consiste en realizar un descortezado, a 75 plantas seleccionadas en lotes de 10 ha donde se busca, por ejemplo, la mielecilla que secreta el insecto y la presencia del insecto. El umbral de acción considerado para comenzar la implementación de las acciones de control se da al detectar un número igual o mayor a 10 piojos por planta (CICESE, 2022).

Con la información recolectada en campo, además de la existencia de bases de datos provenientes de diversas redes de estaciones meteorológicas, y dado que la problemática requiere soluciones y recomendaciones oportunas, es posible aplicar técnicas de aprendizaje de máquina para poder entender la dinámica de la plaga, brindar información, así como recomendaciones para su gestión.

1.3. Hipótesis

- Es posible aplicar técnicas de aprendizaje de máquina basados en ejemplos para ayudar al entendimiento de las condiciones que se tienen que cumplir para que aparezca, perdure o aumente la población del piojo harinoso de la vid.
- La metodología desarrollada puede ser aplicada en otras plagas y otros cultivos.
- Es posible mantener una metodología proactiva basada en datos para la gestión de plagas y enfermedades.

Se propone un flujo de trabajo que parte desde la solicitud, adquisición, fusión, limpieza y transformación de datos para que puedan ser ingresados en un modelo de aprendizaje basado en ejemplos. Este

modelo permite reconocer los patrones existentes entre las características y la variable objetivo (nivel de infestación por PHV) en ranchos vitivinícolas ubicados en el Valle de Guadalupe, principalmente.

La mayoría de las plagas de insectos, así como los hongos y bacterias no regulan su temperatura interna, por lo que su desarrollo depende de la temperatura y condiciones ambientales. Por otro lado, hay insectos como el PHV que requiere personal calificado para su detección, quienes realizan un descortezado e inspección visual rigurosa para lograr detectar la incidencia de la plaga. Dicho esto, procesar información climatológica junto con información de monitoreos constantes sirven para entender las condiciones asociadas con los mayores niveles de infestación. Analizando información existente y colaborando con expertos de diferentes áreas se puede brindar recomendaciones para tener una gestión basada en datos.

Adicionalmente, la metodología seguida para el estudio del PHV en Ensenada, aplicando modelos de aprendizaje basados en ejemplos, se propone la implementación en distintas regiones del país con diferentes plagas y cultivos, es decir, si se cumple la existencia y disponibilidad de datos, es posible entender la dinámica de este tipo de manifestaciones.

Por otra parte, el uso indiscriminado de plaguicidas ha mermado y dañado la actividad agrícola del país. La falta de información principalmente ha impedido adoptar estrategias amigables con el medio que permitan un manejo sustentable de las plagas (Zepeda-Jazo, 2018). Estas estrategias deben considerar los aspectos agroecológicos, económicos y culturales de cada región en particular, partiendo de un análisis entre productores, academia y gobierno con el objetivo de mejorar el equilibrio biológico y la sustentabilidad económica del campo. Sea a pequeña o a gran escala el cultivo depende del desarrollo de nueva tecnología para el manejo de las plagas agrícolas (Zepeda-Jazo, 2018). Actualmente la mayoría de las medidas que se llevan a cabo para el control de diversas plagas y enfermedades que afectan a numerosas unidades de producción se realizan de manera reactiva, es decir, se toman acciones cuando ya existe el problema. Dicho esto, se propone una metodología que pueda ser considerada dentro de las estrategias de manejo integrado de plagas que sea proactiva y que esté fundamentada con toda la información disponible.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar una metodología de apoyo a la toma de decisiones agrícolas que facilite la predicción sobre posibles apariciones de plagas o enfermedades en los cultivos monitoreados.

1.4.2. Objetivos específicos

- Determinar las variables que más influyen en el desarrollo de plagas en el cultivo de estudio.
- Explorar métodos y técnicas que permitan una detección temprana de plagas y enfermedades por medio de modelos predictivos, así como su nivel de fidelidad.
- Determinar cuales son los modelos de aprendizaje de máquina adecuados para esta tarea.
- Determinar si es posible obtener modelos de predicción flexibles que puedan ser empleados para diferentes tipos de cultivos.

Una ventaja que tienen los algoritmos de aprendizaje automático versus los algoritmos de aprendizaje profundo es la explicabilidad. Existen muchos problemas en donde la interpretación de los resultados es crucial, por ejemplo, para elegir el tratamiento médico de un paciente o conocer las condiciones que se cumplieron para la afirmación/negación de cierto fenómeno; básicamente, cualquier decisión que deba estar respaldada por una explicación. Otra ventaja está relacionada al costo computacional; debido a la cantidad de datos que se procesan y a la complejidad de los cálculos matemáticos, los sistemas de aprendizaje profundo requieren un hardware mucho más potente que los sistemas de aprendizaje automático.

Hay muchas aplicaciones donde se pueden aprovechar las ventajas de los métodos tradicionales. Los algoritmos de aprendizaje automático basados en ejemplos son un tipo de algoritmos que en lugar de realizar una generalización explícita, comparan las nuevas instancias del problema con las instancias vistas en el entrenamiento, que han sido almacenadas en la memoria. Se denominan basado en instancias o ejemplos porque construyen las hipótesis directamente a partir de las propias instancias de entrenamiento. Esto significa que la complejidad de la hipótesis puede crecer con los datos, en el peor de los casos, una hipótesis es una lista de n elementos de entrenamiento y la complejidad computacional de clasificar solo una instancia nueva es O(n) (Russell y Norvig, 2009). Una de las ventajas que tiene el aprendizaje basado en instancias sobre otros métodos de aprendizaje automático es su capacidad para adaptar el modelo a datos no vistos previamente, además, pueden simplemente almacenar una nueva instancia o desechar una instancia antigua.

El documento esta ordenado de la siguiente manera. En el capítulo 2 se muestran los métodos y materiales necesarios para lograr el objetivo planteado, se muestra el trabajo relacionado, el caso de estudio correspondiente con el Piojo Harinoso de la Vid (PHV) describiendo los datos que se emplearon. En

seguida se presentan los modelos de aprendizaje con los que se trabajó. En el capítulo 3 se expone la metodología que se llevó a cabo, partiendo desde la recolección y limpieza de datos, la selección de los predictores, así como los diversos modelos de aprendizaje utilizados. A continuación se muestran los resultados obtenidos en el capítulo 4, así como la discusión de los mismos y se añade una serie de recomendaciones. Se finaliza con unas conclusiones del trabajo, así como trabajo futuro.

Capítulo 2. Materiales y Métodos

En este capítulo se detalla el trabajo relacionado y la literatura revisada para llevar a cabo el objetivo planteado. Se describen además los datos empleados y los modelos seleccionados para realizar las predicciones, los métodos de selección de características, así como las métricas con las cuales se evaluaron los resultados.

2.1. Predicción de Plagas y Enfermedades

Actualmente las técnicas de aprendizaje de máquina (ML), tanto somero como profundo, son líderes en varios campos, por ejemplo, aplicaciones médicas, reconocimiento de imágenes, procesamiento de audio y voz, entre otras áreas. La gran cantidad de datos que se genera día con día hace posible análisis masivos de información. La aplicación del ML en sectores primordiales como lo es la agricultura presenta diversas aplicaciones dentro de las que resaltan mejor gestión del cultivo y de los recursos.

La agricultura 4.0 ofrece muchas posibilidades, por ejemplo, al emplear drones y redes de sensores se puede proporcionar información, en tiempo real o diferido, de parámetros agronómicos para alertar a los agricultores sobre el progreso de un cultivo, el estado del suelo, el crecimiento de maleza y la aparición o riesgo de plagas y enfermedades (Santos Valle y Kienzle, 2020). Las plagas y enfermedades de las plantas pueden propagarse fácilmente a varias regiones, alcanzar dimensiones de epidemia y provocar innumerables pérdidas, poniendo en peligro los medios de vida de los agricultores vulnerables, así como la seguridad alimentaria y nutricional de millones de personas (FAO, 2021).

A continuación se presenta un análisis de publicaciones que abordan el tema de predicción de plagas y enfermedades en diversos cultivos mediante técnicas de aprendizaje de máquina, basándose principalmente en parámetros climatológicos así como de información de vigilancia de los cultivos y de las plagas o enfermedades.

Durgabai et al. (2018) expone los beneficios que se tienen al aplicar técnicas de aprendizaje automático para detectar plagas en fases iniciales. El ML ayuda a descubrir reglas y patrones en los conjuntos de datos ya que es posible tomar en cuenta muchos de los posibles factores: datos históricos, datos de satélites, datos de sensores, información climática, información del suelo, etc. Al aplicar el aprendizaje automático a los datos, los sistemas de gestión agrícola se convierten en programas habilitados en tiempo real, que proporcionan recomendaciones e información para la toma de decisiones y acciones de los agricultores (Liakos et al., 2018).

Chakraborty et al. (2004) recopilaron y analizaron datos de las condiciones meteorológicas y de la gravedad de la enfermedad antracnosis que afecta seriamente a las leguminosas y a otras plantas de siete campos de Australia, Brasil y Colombia. Los datos de gravedad de la enfermedad y condiciones meteorológicas de uno o varios campos fueron analizados mediante redes neuronales artificiales con el objeto de predecir la enfermedad de la gravedad en otro lugar. El error global obtenido fue de $21.9\,\%$ y las variables meteorológicas más importantes fueron las relacionadas con la humedad, lluvia, humedad de la hoja, radiación y viento.

Kaundal et al. (2006) muestran un caso de estudio para predecir la enfermedad de tizón del arroz (*rice blast*) en India. Emplearon datos recolectados durante cinco años en un ambiente controlado. Las variables climatológicas de mayor influencia fueron lluvia, humedad relativa máxima y mínima, temperatura máxima y mínima . El modelo que mejor resultados mostró fue Máquina de Soporte Vectorial (SVM), incluso comparando contra redes neuronales artificiales.

Klem et al. (2007), Wang y Ma (2011) y otros, se interesaron por analizar plagas y enfermedades que afectan a uno de los cereales más importantes de la alimentación; el trigo. Klem et al. (2007) analizaron datos sobre el contenido de la toxina DON (deoxinivalenol) en granos de trigo, las condiciones meteorológicas y las prácticas de cultivo que se llevaron a cabo en dos experimentos de campo realizados en 2005. Con esta información se desarrolló un modelo de redes neuronales artificiales para predecir el contenido de DON; los parámetros de entrada para realizar la predicción fueron la cosecha anterior, dos temperaturas promedio y las precipitaciones. Por otro lado, Wang y Ma (2011) analizaron la enfermedad de la roya de trigo por medio de máquinas de soporte vectorial (SVM) para predecir la enfermedad y elaborar estrategias de control para cultivos de China. La precisión reportada va desde $50\,\%$ hasta $100\,\%$ de correctas clasificaciones, esto considerando distintas configuraciones del modelo y de los conjuntos de datos.

Hay estudios que parten desde la instalación de redes de sensores para el objetivo particular de realizar predicciones de plagas y enfermedades, tal es el caso de Tripathy et al. (2011) quienes llevaron a cabo un experimento en una región semiárida de India para comprender las relaciones cultivo-clima-plaga/enfermedad utilizando datos sensoriales inalámbricos y vigilancia a nivel de campo sobre la dinámica de las plagas (Trips) y enfermedades (Necrosis), afectaciones estrechamente relacionadas con cultivos de cacahuate. Utilizaron técnicas de extracción de datos para obtener tendencias y correlaciones entre cultivo-clima-plaga/enfermedad. Empleando regresión multivariable, reglas de asociación y redes bayesianas obtuvieron modelos que asocian las variables predictoras (p.ej., Temperatura máxima, humedad relativa y evotranspiración) con las variables objetivo (porcentaje de incidencia de la plaga o enfermedad).

Pérez-Ariza et al. (2012) demostraron la efectividad de emplear redes bayesianas para predecir la incidencia de la enfermedad de la roya en café, enfermedad que produce hasta un $35\,\%$ de pérdidas. Los datos fueron recolectados en una granja experimental en Brasil durante ocho años. De manera similar la temperatura y humedad, así como la lluvia, fueron las variables con mayor impacto, además de otras variables como son el mes y la temporada del año, fueron determinadas de relevancia.

En Xiao et al. (2019) emplearon el algoritmo a priori para encontrar reglas de asociación entre los factores climáticos y la ocurrencia de diversas plagas y enfermedades que afectan a los cultivos de algodón a lo largo de seis regiones de la India. Posteriormente emplearon redes LSTM (*Long Short Term Memory*) para realizar las predicciones de presencia o ausencia de las afectaciones analizadas, la precisión alcanzada fue de 0.87 a 0.92. Otro estudio se muestra en Aparecido et al. (2019), quienes utilizaron diversos modelos de regresión para predecir el porcentaje de incidencia de la roya del café, la cercospora, el minador y perforador del café, usando datos meteorológicos y de campo en regiones de Brasil. El modelo con mayor precisión fue el Random Forest Regressor, y las variables de mayor importancia fueron temperatura máxima y humedades relativas.

La publicación de Rodríguez-Moreno et al. (2020) fue realizada con datos nacionales del estado de Sonora. Se interesaron por la enfermedad fúngica de mayor importancia económica en cultivos de cereales del mundo; la roya del trigo de invierno. Emplearon árboles de clasificación y regresión (CARTs) para predecir la presencia/ausencia de patógenos con información meteorológica, así como de incidencia y severidad de la roya de trigo.

De manera resumida y para fines comparativos, en la tabla 2 se muestra el país, el cultivo, la plaga o enfermedad, los modelos de ML y los parámetros empleados por Xiao et al. (2019), Aparecido et al. (2019) y Rodríguez-Moreno et al. (2020).

Tabla 2. Comparación de tres publicaciones.

Publicación	Xiao et al. (2019)	Aparecido et al. (2019)	Rodríguez-Moreno et al. (2020)
País	India	Brasil	México
Cultivo	**Algodón**	**Café**	**Trigo**
Plaga o enfermedad	Gusano cogollero, mosca blanca saltahojas, tizón de la hoja	Roya del café y Cercospora Minador de hojas, Barrenador de bayas	Roya de la hoja de trigo Roya amarilla del trigo
Método	Clasificación: LSTM (DL)	Regresión: kNN, RFR, RLM	CART's
Resultado	Presencia o Ausencia	% de incidencia	Incidencia (n) y severidad(%)
Parámetros	Temp Max (°C), Temp Min (°C), Humedad Rel en la mañana (%), Humedad Rel en la noche (%), Luvias (mm), Velocidad del Viento (kmph), Horas sol (hrs) Evaporatión (mm).	Temp Min (°C), Temp Max (°C), Lluvia total (mm), NDR $\geq$ 1 mm NDR $\geq$ 10 mm Humedad Relativa(%), NdRH90 % , NdRH80 %	Temperatura (°C), Lluvia (mm), Humedad Relativa (%), Temperatura de punto de rocío (°C), Máxima Velocidad del Viento (km.h-1), Dirección del Viento (°Azimuth), Temperatura en la noche (°C), Radiación Solar (W.m-2).

Por último, y para un cultivo de interés como lo es la *Vitis Vinífera*, se tiene el estudio de Chen et al. (2020), quienes emplearon datos climáticos y de campo recolectados durante 9 años para predecir la incidencia y severidad de una enfermedad fúngica conocida como Cenicilla vellosa (*Grape Downy Mildew*), una de las enfermedades más dañinas para los cultivos de vid. Las predicciones fueron llevadas a cabo con diversos métodos, y tras analizar la sensibilidad y especificidad, aunado a añadir la fecha de inicio de la enfermedad como predictor un modelo de ensamble (*gradient boosting)*, obtuvieron los mejores resultados, con valores de AUC (área bajo la curva) desde 0.65 hasta 0.86.

Con base en la revisión de literatura se mencionan ciertas observaciones, así como algunas de las contribuciones de la tesis.

- La tarea de predicción empleando parámetros meteorológicos se ve reducida si se compara con publicaciones que emplean imágenes y aprendizaje profundo.
    - El trabajo pretende dar recomendaciones previas, es decir, se busca una metodología proactiva basada en datos existentes para la gestión de plagas y enfermedades, además, la recolección de fotografías con su correspondiente etiquetado esta fuera de los alcances del proyecto.
- Las publicaciones mostradas enfatizan cuáles son las variables más importantes para que aparezca o aumente la plaga o enfermedad en cuestión.
    - Esto es uno de los principales objetivos del presente estudio.
- La mayoría de los datos empleados en las publicaciones son recolectados en granjas experimentales, o bien, son conjuntos de datos disponibles de publicaciones previas.
    - En el presente proyecto de tesis se trabaja con una base de datos recolectada bajo metodologías establecidas por especialistas en agricultura con el objetivo de realizar un seguimiento de la problemática y dar sugerencias de control. Se trabaja con información no pública, la cual no tenía como objetivo ser empleada para una aplicación de ML.
- Todas las publicaciones se enfocan en un cultivo determinado.
    - Si bien la tesis se enfoca en cultivos de vid se logró generar un convenio en donde se analizará un segundo caso de estudio con cultivo y plaga diferente.

2.2. Caso de Estudio

El estado de Baja California es uno de los principales estados con mayor producción de vid junto con Sonora y Zacatecas. En particular, la uva industrial es de gran importancia en la región pues el $70\,\%$ de vino nacional proviene de Baja California (SADER, 2018). El caso de estudio seleccionado nace del interés en resolver un problema fitosanitario que ocurre en distintas regiones vinícolas del mundo y que actualmente presenta afectaciones en territorio nacional.

2.2.1. Piojo Harinoso de la Vid (PHV)

Existen diferentes especies de piojos harinosos, los cuales se relacionan con enfermedades virales, pero la especie de mayor importancia en la región es el *Planococcus ficus* cuya variedad proviene de Egipto e Israel y del sur de Europa. *Planococcus ficus* se registró por primera vez en 1994 en California y en el Valle de Coachella en 1998, EE.UU. En México se detectó por primera vez en noviembre de 2000 en el estado de Sonora en un viñedo de vid para mesa de 150 hectáreas, donde el $100\,\%$ de la producción fue afectada, provocando pérdidas mayores a los 2 millones de dólares (Castillo et al., 2004), y en el estado de Baja California se registró en el año 2014 en el Valle de Guadalupe (Agricultura-Senasica, 2021). Los piojos harinosos también afectan a varios cultivos como son aguacate, dátil, higo, manzana, naranja y plátano, aunque son diferentes especies.

El PHV es una plaga clave en las regiones vitivinícolas más importantes del mundo, tales como Argentina, California, Europa, África mediterránea, México, etc. El reciente establecimiento y propagación del PHV en Brasil confirma que este piojo es una especie invasora peligrosa en las zonas vitivinícolas de todo el mundo (Cocco et al., 2021). De acuerdo con la Fundación Americana de Viñedos (AVF) esta plaga se encuentra entre los principales insectos de la vid, es la principal preocupación de los viticultores y presenta una amenaza para la sostenibilidad de la industria del vino y de acuerdo con Sharma et al. (2018), se requiere una investigación de alta prioridad.

Entre algunos de los daños directos e indirectos que provoca esta plaga se encuentra: el agotamiento de la savia, caída temprana de hojas, muerte regresiva, debilitamiento, disminución de la calidad de los vinos, en la melaza se pueden desarrollar hongos de fumagina, es vector del virus asociado con enfermedades de enrollamiento de la hoja y mancha roja (GLRV, RBV). También afecta a la calidad estética, reduciendo su valor en el mercado y es causa de rechazo para exportación. Las pérdidas se estiman hasta un $30\,\%$ de la producción. Sin embargo, cuando no se hace un manejo o control adecuado se pueden tener pérdidas del $100\,\%$ (Castillo et al., 2004). El PHV tiene una alta tasa de reproducción; una hembra en promedio

pone 250 huevos y se tienen de 4 a 6 generaciones por año, también hay generaciones superpuestas, es decir, que se encuentran todos los estados de desarrollo en la misma temporada (CICESE, 2022).

Figura 5. Infestación de piojo harinoso, fotografías tomadas en campo.

Los programas convencionales de control de PHV en viñedos se basan en aplicaciones repetidas de insecticidas sintéticos a lo largo de la temporada de crecimiento de la uva, incluidos los organofosforados y los neonicotinoides, que tienen efectos graves en las abejas y polinizadores. La eficacia del control químico es variable y a menudo insatisfactorio, ya que los piojos residen principalmente en lugares ocultos; debajo y en grietas de la corteza así como en las raíces hasta 30 cm de profundidad (Cocco et al., 2021).

En los cultivos de Ensenada las medidas de control se dividen en control cultural, físico, biológico y químico. El control cultural es la acción que implica descortezado de la planta para exponer al piojo al clima o contacto con los productos químicos, es una forma de control muy efectiva pero muy costosa. El control físico consiste en tratamientos hidrotérmicos durante 5 minutos a 51^oC, además de que mata al PHV, también sirve como tratamiento para plagas como nemátodos agalladores, filoxera y patógenos bacterianos, pero esto solo se puede hacer en esquejes (Agricultura-Senasica, 2021). También se han implementado tratamientos de interrupción del apareamiento mediante feromonas de confusión sexual. En términos de control biológico se han reportado el uso de parasotoides y depredadores en California

y en Sonora (Joyce et al., 2001; Castillo et al., 2004). Sin embargo, una de las principales causas por las que los enemigos naturales no realizan eficientemente su acción de control es por la presencia de hormigas con las que convive el PHV, ante esto, se tienen que realizar acciones de control de las hormigas previo a la liberación de los agentes de control biológico (Agricultura-Senasica, 2021). Finalmente, el control químico es la acción que se realiza mayoritariamente, emplea productos químicos donde los más utilizados en México son imidacloprid y spirotetramat con reducciones de hasta $73\,\%$ de la población. También hay plaguicidas botánicos y biológicos, como lo es el aceite de neem, el cual ayuda a detener la reproducción del piojo (CICESE, 2022).

2.2.2. Datos

Información de campo

Una de las acciones que hace el grupo de técnicos del Comité Estatal de Sanidad Vegetal de Baja California (CESVBC) es realizar muestreos desde el año 2017. Abarcan un área de 3,800 hectáreas, atendiendo a 214 predios propiedad de 167 productores aproximadamente. Los muestreos se realizan en lotes de 10 hectáreas, en cada lote se recorren tres líneas de muestreo y se revisan 25 plantas por línea seleccionadas al azar acumulando 75 plantas en total; cada una se descorteza tomando varios puntos desde la base del tronco hasta los cordones, y se registran el número de piojos encontrados. Con la información del muestreo se determina el nivel de infestación de cada lote de acuerdo con la tabla 3.

Tabla 3. Niveles de infestación.

	Sin presencia	**Nivel Leve**	**Nivel Medio**	**Nivel Fuerte**
% Plantas Infestadas	0	1-20	21-30	30
Piojo/Planta	0	1-30	30-100	100

Además de indicar el porcentaje de plantas infestadas y del número de piojos por planta, el personal añade información correspondiente a la ubicación, la fecha, la variedad de la uva, la superficie muestreada, el tipo de daño encontrado, la fenología, etc. La colección de esta información es almacenada anualmente en las oficinas pertinentes, y gracias a la vinculación y colaboración del comité de este proyecto con los interesados que desean encontrar soluciones al problema, se generó un convenio en el cual se tuvo acceso a los datos recolectados para poder ser procesados considerando las medidas de confidencialidad necesarias.

Dentro de las variables registradas existe una de particular interés: la fenología. La fenología define el crecimiento y el desarrollo que están relacionados con factores climáticos y que se repiten periódicamente

en plantas y animales; también estudia los estadíos de crecimiento de los seres vivos y sus relaciones con diferentes factores climáticos. En las plantas la brotación, la expansión de las hojas, la floración, la formación de semillas, la fructificación, la dispersión de semillas y la germinación tienen lugar a su debido tiempo. La fenología de la *vitis vinífera* varía dependiendo la fuente. Por ejemplo, Valladolid et al. (2018) describe 17 estados fenológicos, sin embargo, se pueden distinguir cuatro estados fenológicos principales: brotación, floración, cuajado y maduración. En los muestreos realizados en campo por parte el grupo del CESVBC se identifican y registran las siguientes etapas:

- Brotación: Esta etapa consiste en la aparición de la punta verde constituida por el brote joven, después de haberse activado el desarrollo de la yema en reposo. En esta etapa se da la coincidencia total de capacidad de desbarrar de las yemas sin ningún tipo de inhibición con las adecuadas condiciones climatológicas (Valladolid et al., 2018).

- Floración: Momento en el que aparecen los embriones de las flores, que posteriormente darán lugar a los granos de las uvas que formarán los racimos. Este proceso se desarrolla en condiciones normales en el mes de junio. Se considera que el viñedo está en plena floración cuando este fenómeno ha ocurrido en al menos el $50\,\%$ de las inflorescencias, y se considera final de la floración cuando ha tenido lugar en todas las flores de las inflorescencias de todas las cepas del viñedo (Valladolid et al., 2018).

- Crecimiento de baya: Este proceso comienza con la caída de las flores no fecundadas o los frutos mal cuajados. En los cultivos con semilla, el crecimiento de la baya tiene lugar a partir de la polinización y fecundación, aproximadamente durante las 2-3 semanas siguientes a la floración.

- Post-cosecha: Etapa luego de un ciclo de producción intenso, la planta sufre un desgaste severo de nutrientes y de energía. Es el periodo de tiempo que transcurre entre cosecha y caída de hojas.

- Reposo: Durante los meses del invierno las yemas se encuentran en dormancia y no existe indicio de crecimiento. En este estado no hay crecimiento aparente.

- Poda: La poda consiste en la supresión total o parcial de sarmientos, hojas, racimos u otras partes vegetativas de la planta. Usualmente se refiere a la poda como a la práctica de eliminación de sarmientos en el periodo de receso invernal (Valladolid et al., 2018).

Figura 6. Etapas fenológicas de la vid.

Información meteorológica

Actualmente existen múltiples redes de estaciones meteorológicas compuestas de un conjunto de sensores de medición que se encuentran distribuidas en todo el territorio nacional, p.ej., Sistema de Información para el Manejo de Agua de Riego en Baja California SIMARBC, Red de estaciones Meteorológicas de Ensenada CICESE, Estaciones Metereológicas Automáticas de la CONAGUA, entre otras. Estas estaciones registran constantemente múltiples parámetros meteorológicos (temperatura, humedad relativa, precipitación, dirección y velocidad del viento, radiación solar etc.) por medio de los diversos sensores con los que cuentan.

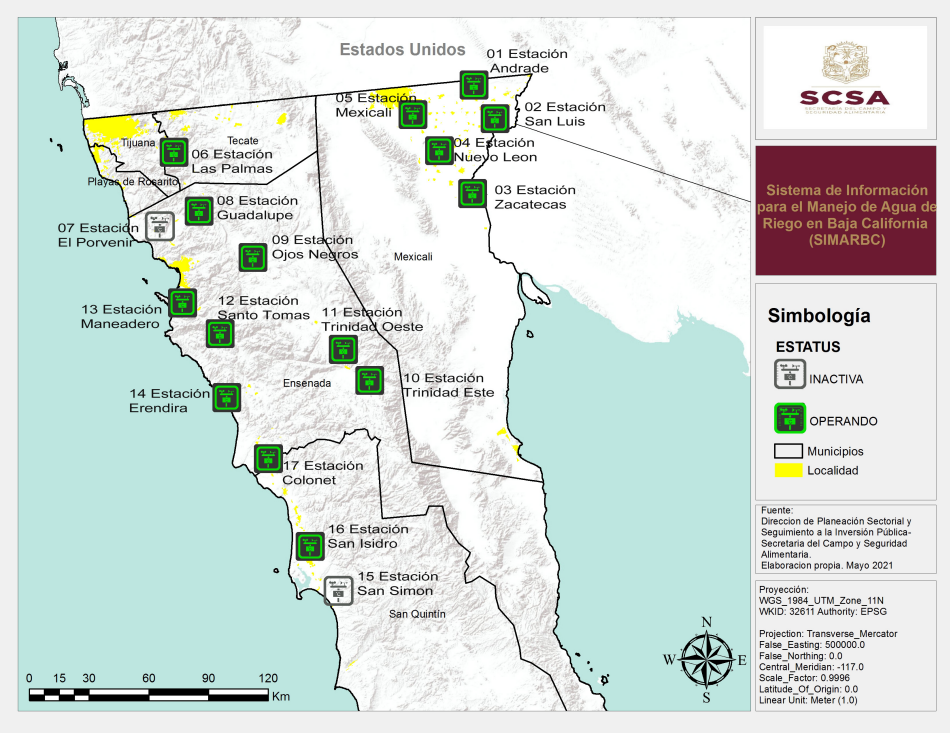

Figura 7. Ubicación geográfica de las estaciones del SIMARBC.

Dada la disponibilidad y distribución geográfica de las estaciones, se optó por trabajar con la red de

estaciones SIMARBC, la cual consiste de 17 estaciones distribuidas en el estado de Baja California (ver Fig. 7). De cada estación se puede acceder a una gran cantidad de variables meteorológicas y agroclimáticas, las cuales se muestran en la tabla 4, sea por hora o por día, desde el año 2005 hasta la fecha actual.

Tabla 4. Variables registradas por las estaciones del SIMARBC.

Variable	Unidad de medida
Evapotranspiración (Eto)	mm
Precipitación	mm
Radiación solar	Cal/cm^2
Presión vapor	Kpas
Temperatura máxima	°C
Temperatura mínima	°C
Temperatura aire	°C
Temperatura promedio	°C
Humedad relativa	%
Humedad relativa promedio	%
Humedad relativa máxima	%
Humedad relativa mínima	%
Punto de rocío	°C
Punto de rocío promedio	°C
Punto de rocío máxima	°C
Temperatura del suelo	°C
Temperatura promedio del suelo	°C
Velocidad del viento	m/s
Dirección del viento	°N
Unidades calor	U.C
Horas frío (Mexicali)	H.F
Horas frío (Costa)	H.F

2.3. Modelos de aprendizaje de máquina

El aprendizaje de máquina, también llamado aprendizaje automático o ML, se divide principalmente en aprendizaje supervisado el cual necesita conjuntos de datos etiquetados, para realizar tareas de clasificación y regresión. La etiqueta para un modelo de clasificación es de tipo discreta dentro de un conjunto limitado de etiquetas (llamadas clases también) mientras que la etiqueta para regresión es un número real (entero o complejo, vector, etc.). Por otro lado, cuando no disponemos de etiquetas se pueden aplicar técnicas de aprendizaje no supervisado; este tipo de modelos tienen el objetivo de comprender y abstraer los patrones de la información directamente. Adicionalmente se tiene el aprendizaje por refuerzo; en esta técnica los modelos aprenden a partir de la experiencia. En la figura 8 se muestran ejemplos de aplicación de los tipos de aprendizaje de máquina mencionados.

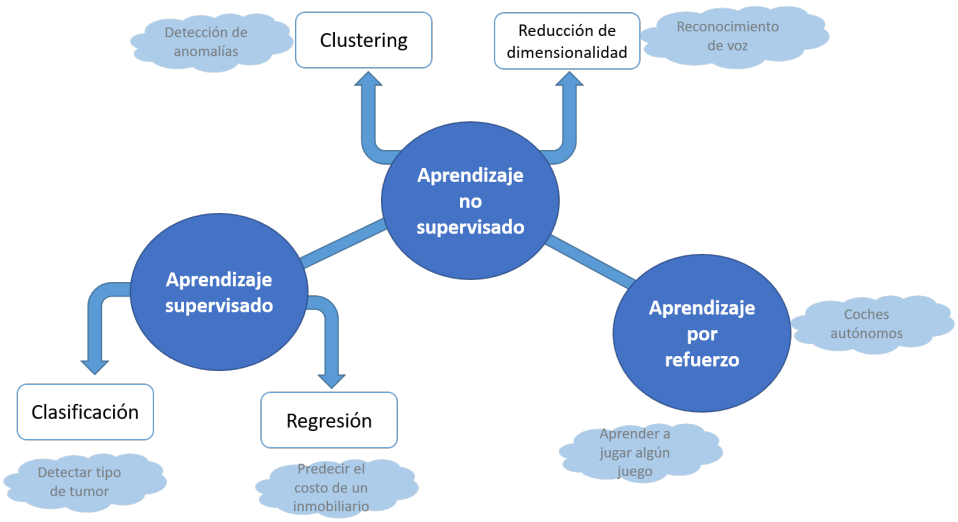

Figura 8. Tipos clásicos de aprendizaje de máquina, con algunos ejemplos de aplicación.

Los modelos de aprendizaje basados en ejemplo, o *Instance-based learning*, son de tipo aprendizaje supervisado, los cuales emplean los ejemplos almacenados en una base de datos para compararlos directamente con las nuevas consultas. Son basados en instancias y utilizan un modelo implícito, en donde se almacena la información y el procesamiento se realiza en el momento de una nueva consulta. Dentro de este tipo de algoritmos es donde se ubica el algoritmo kNN y el HSP, modelos que fueron empleados en el presente estudio.

2.3.1. K Vecinos Cercanos (kNN)

El clasificador de K vecinos cercanos (kNN) calcula la distancia entre cada consulta y cada uno de los ejemplos dentro del conjunto de entrenamiento, con lo cual se obtiene la vecindad conformada por los k ejemplos u objetos más cercanos. La asignación se realiza generalmente considerando la clase más repetida en el vecindario. En la figura 9 se muestra un ejemplo representativo seleccionando un valor de $k = 3$ para un problema de clasificación binaria con datos de dos dimensiones, es decir, dos variables o características (x_1 y x_2). La consulta q_1 se asigna directamente a la clase 1, pues sus k vecinos son de esta clase, mientras que para la consulta q_2 es necesario aplicar una regla para decidir a qué clase asignar, en este caso con una regla de mayoría simple es suficiente, la cual asigna la consulta en la clase 2.

El algoritmo kNN requiere de una medida de distancia que permita comparar los ejemplos almacenados con las nuevas consultas que se desean clasificar. En la mayoría de los casos los conjuntos de datos son representados por vectores en espacios multidimensionales, con un conjunto de entrenamiento X

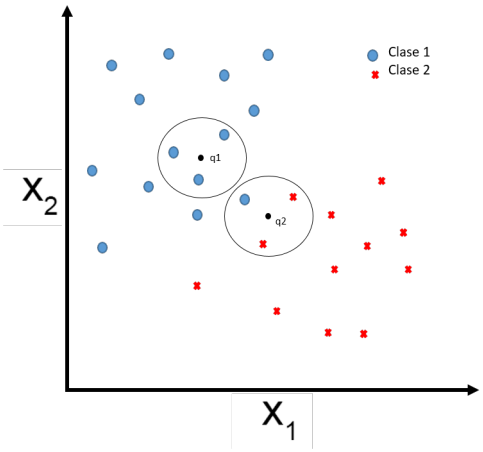

Figura 9. kNN con k=3 en espacio de dos dimensiones.

conformado por $\{x_i\}i \in [1, n]$ ejemplos de entrenamiento (con n= cantidad de datos). Los vectores con los ejemplos de entrenamiento están descritos por un conjunto F características o variables que corresponden con el número de dimensiones, es decir, un ejemplo con 10 características corresponde a un espacio de 10 dimensiones. Como se trata de aprendizaje supervisado, cada ejemplo de entrenamiento está clasificado con una etiqueta $y_j \in Y$. El objetivo es clasificar cada ejemplo desconocido q. Para cada $x_i \in X$ se calcula la distancia entre q y x_i:

$$d(q, x_i) = \sum_{f \in F} w_f \delta(q_f, x_{if}) \tag{1}$$

Con la ecuación 1 se tiene una sumatoria sobre todas las F características con peso w_f para cada característica. Se pueden utilizar muchas métricas, la más común es la distancia Euclideana, pero se pueden aplicar otras como la distancia de Manhattan, Chebyshev, Mahalanobis, etc.

El desempeño de kNN depende en gran medida del tamaño del vecindario que se define por el parámetro k. Elegir el valor óptimo de k no es una tarea sencilla; propuestas diversas se han generado para escogerlo automáticamente. En el estado del arte se encuentran los métodos *kTree* y *k*Tree*, los cuales agregan una etapa previa al entrenamiento del algoritmo kNN para encontrar los valores óptimos de k para cada consulta (Zhang et al., 2017). En efecto, la principal debilidad del clasificador kNN es la necesidad de establecer el valor del hiperparámetro k, el cual es crítico para obtener buenos resultados en la clasificación, tal como se representa en la figura 10.

Con la métrica de distancia y el valor de k especificados se necesita una regla para que, con base en los vecinos seleccionados, se elija la clase de cada consulta. Cuando se trata de clasificación, el algoritmo básico de kNN utiliza una regla de mayoría simple, es decir, se le asigna a la consulta la clase más

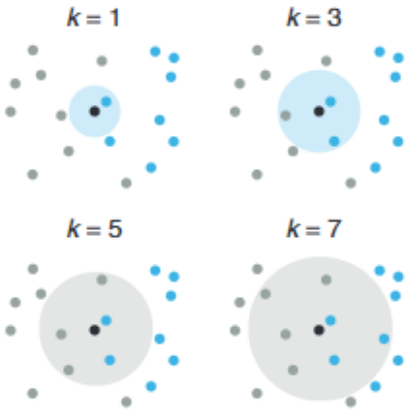

Figura 10. Asignación de un punto no clasificado (negro) basado en voto mayoritario de los k vecinos cercanos del conjunto de entrenamiento (puntos grises y azules). Se muestran los casos para k=1, 3, 5 y 7; los k vecinos están dentro del círculo coloreado por el voto mayoritario de la clase.

repetida entre los vecinos. Sin embargo, existen diferentes reglas que se pueden ocupar para hacer la votación, por ejemplo, hay métodos que le asignan mayor peso a los vecinos que se encuentran más cerca de la consulta.

2.3.2. HSP

Talamantes y Chavez (2022) proponen utilizar el grafo HSP para aplicaciones de aprendizaje basado en ejemplos conservando la simplicidad del algoritmo kNN. En este caso, en lugar de comparar los objetos más cercanos a una consulta, se comparan los objetos obtenidos por el algoritmo del grafo HSP (ver Fig. 11) y después se ejecuta una regla de votación. En esta propuesta se supone que cada consulta puede elegir sus vecinos en la base de datos o en el conjunto de entrenamiento.

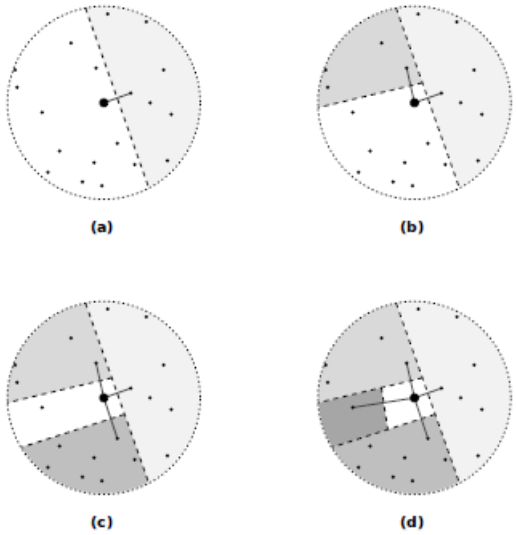

Figura 11. Selección de vecinos para nodo arbitrario.

El área sombreada en la Figura 11 representa los semiespacios que se agregan iterativamente a la región prohibida. Computacionalmente, una arista (μ, z) está prohibida por una arista (u, ν) cuando la distancia euclidiana de z a ν es menor que la distancia euclidiana de z a μ. Es importante resaltar que cada nodo elige a sus vecinos sin necesidad de parámetros.

Los vecinos seleccionados tienen la propiedad de ser similares a la consulta y diversos entre sí, lo que da una vecindad representativa para cada consulta, lo cual es deseable en tareas de aprendizaje basado en ejemplos (Talamantes A., 2021). Además, tiene la ventaja de ser libre de parámetros, evitando así el principal problema de kNN al seleccionar el valor de k. Con este modelo se puede aplicar cualquier métrica de distancia para cualquier dimensión, así como la regla de votación que se considere mejor, según sea el caso. El pseudocódigo del modelo de clasificación HSP se muestra en la figura 12

```
   Entrada: conjunto de entrenamiento X y conjunto de prueba Y
   Salida: etiquetas de Y
 1 para cada u ∈ Y hacer
 2 |   N ← ∅;
 3 |   C ← X;
 4 |   mientras C sea no vacío hacer
 5 |   |   v ← c ∈ C | d(u, c) ≤ d(u, c'), ∀c' ∈ C;
 6 |   |   N.insertar(v);
 7 |   |   para cada c ∈ C hacer
 8 |   |   |   si d(c, u) > d(c, v) entonces
 9 |   |   |   |   C.remover(c);
10 |   |   |   fin
11 |   |   fin
12 |   fin
13 |   etiqueta(u) ← etiqueta más repetida en N;
14 fin
```

Figura 12. Algoritmo para clasificación con HSP.

2.3.3. Selección de Características

La Selección de Características (*Feature Selection FS*) es el proceso de seleccionar las características más importantes o relevantes de un conjunto de datos con el objetivo de mejorar el rendimiento de predicción. Este proceso selecciona un subconjunto de características claves de los datos originales en un intento por disminuir la dimensionalidad del problema. Con la FS se eliminan características irrelevantes, redundantes o altamente correlacionadas.

La selección de características se puede usar para tareas de clasificación o regresión. El subconjunto de características seleccionado debe representar la máxima varianza de los datos y este subconjunto de características se utiliza para entrenar el modelo. La FS tiene dos propósitos principales: en primer lugar,

la selección de características a menudo aumenta la precisión de la clasificación a través de la eliminación de características irrelevantes, redundantes o altamente correlacionadas; en segundo lugar, disminuye la cantidad de características, lo que hace que el proceso de entrenamiento del modelo sea más eficiente.

A pesar de que la selección de características sí disminuye la cantidad de características en el conjunto de datos que se usa para entrenar el modelo, no es lo mismo que el término reducción de la dimensionalidad. Los métodos de selección de características extraen un subconjunto de las características originales de los datos sin cambiarlas, mientras que los métodos de reducción de dimensionalidad emplean características diseñadas que pueden transformar las características originales y, de ese modo, modificarlas.

De manera general hay dos tipos de técnicas de selección de características: supervisada y no supervisada. Dentro de la selección supervisada se identifican tres principales métodos; *filter, wrapper* y *embedded* (Fig. 13).

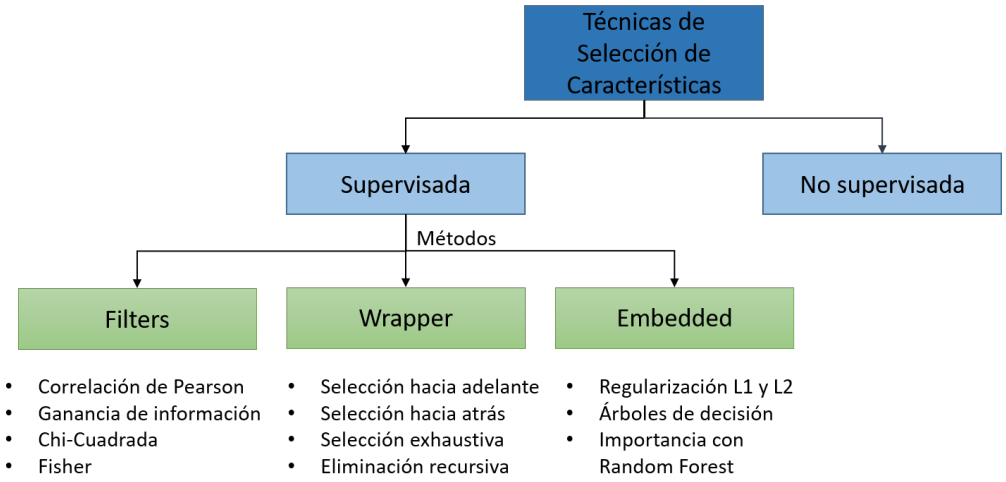

Figura 13. Técnicas de selección de características.

Los métodos de selección de características basados en filtros aplican una medida estadística para asignar una puntuación a cada característica. En seguida, las características se clasifican según la puntuación, esto se puede usar para definir el umbral para conservar o eliminar una característica específica. Algunos ejemplos de estas medidas estadísticas son el coeficiente de correlación de Pearson, prueba de Chi-cuadrada, ganancia de información, prueba de Fisher, ANOVA, etc.

En los métodos wrapper, la selección de características se realiza considerando un problema de búsqueda, se realizan diferentes combinaciones, se evalúan y se comparan con otras. Se entrena el algoritmo utilizando un subconjunto de características de forma iterativa. Sobre la base del resultado del modelo, se añaden o restan características y el modelo se entrena de nuevo. Algunas técnicas son: selección

hacia adelante, eliminación hacia atrás, selección exhaustiva de características y eliminación recursiva de características.

Finalmente, los métodos embebidos combinan las ventajas de los métodos de filtro y los wrappers al considerar la interacción de las características junto con un bajo costo computacional. Estos métodos también son iterativos, evalúan cada iteración y encuentran de forma óptima las características más importantes que más contribuyen al entrenamiento de una iteración concreta. Algunas técnicas estos métodos integrados son regularización y los métodos basados en árboles.

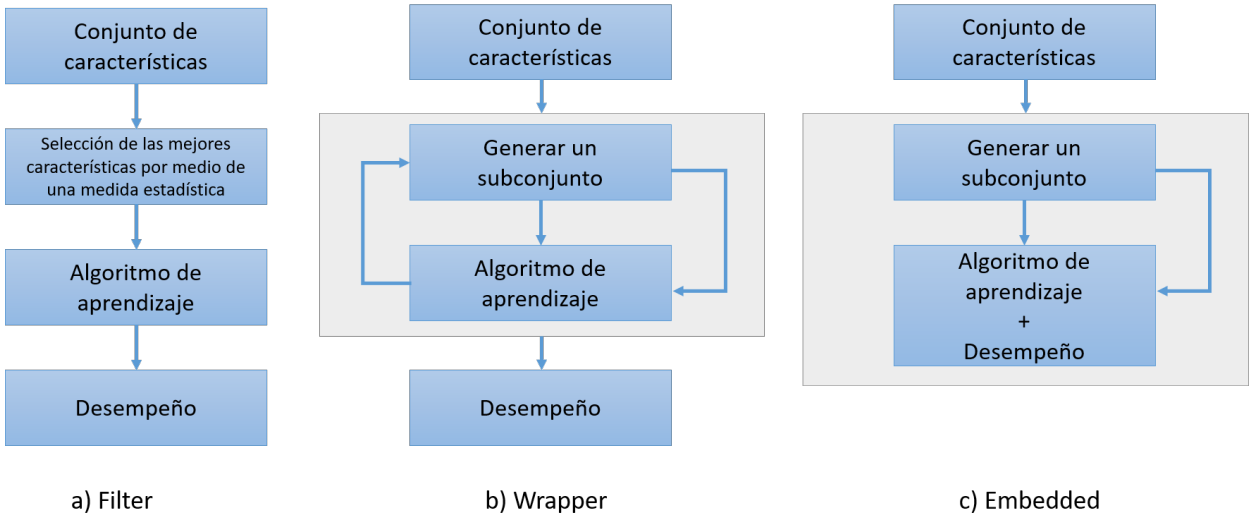

Figura 14. Funcionamiento de los métodos de selección de características.

2.3.4. Métricas

Como ya se mencionó, los modelos de aprendizaje empleados para clasificación *aprenden* de un conjunto de datos de entrenamiento y el modelo entrenado se prueba con un conjunto de datos no visto (conjunto de prueba) de forma experimental. El rendimiento experimental que se obtiene con los datos de prueba es un indicador del desempeño y comprueba la capacidad de generalización del clasificador. En esta sección se muestran algunas métricas que se emplean para evaluar el desempeño de los modelos de clasificación.

- Accuracy: Es el porcentaje de clasificaciones correctas.
- Tasa de error: Porcentaje de clasificaciones incorrectas.

Estas métricas sencillas presentan algunos problemas: asumen costos iguales para las clasificaciones erróneas y asumen una distribución de clases relativamente uniforme. Ante esto, otras métricas resultan más representativas, las cuales se pueden obtener de la matriz de confusión.

Matriz de confusión

Una matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. En términos prácticos nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo.

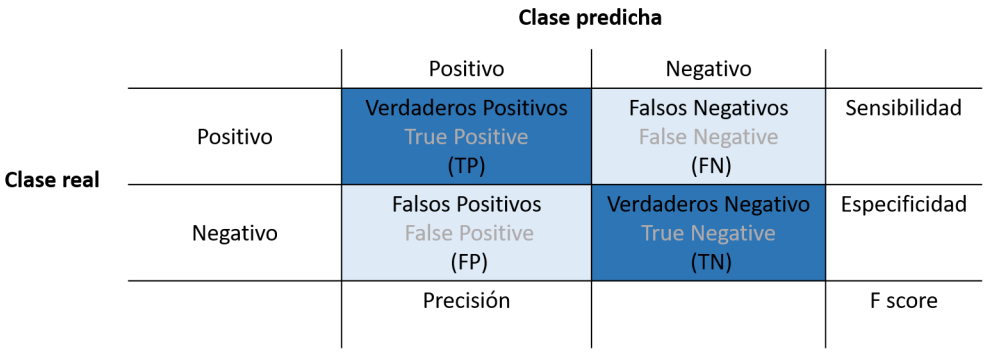

Figura 15. Matriz de confusión para una clasificación binaria.

En una matriz de confusión binaria, como la que se observa en la figura 15, se tienen cuatro posibles resultados. Para explicar cada posibilidad se emplea un ejemplo de pruebas para detectar el covid:

- Verdadero Positivo (TP): El valor real es positivo y la prueba predijo también que era positivo, *i.e.*, una persona está enferma y la prueba así lo demuestra.
- Verdadero Negativo (TN): El valor real es negativo y la prueba predijo también que el resultado era negativo, *i.e.*, la persona no está enferma y la prueba así lo demuestra.
- Falso Positivo (FP): El valor real es negativo, y la prueba predijo que el resultado es positivo, *i.e.*, la persona no está enferma, pero la prueba nos dice de manera incorrecta que si lo está. Esto es lo que en estadística se conoce como error tipo I
- Falso Negativo (FN): El valor real es positivo, y la prueba predijo que el resultado es negativo, *i.e.*, la persona está enferma, pero la prueba dice de manera incorrecta que no lo está. Esto es lo que en estadística se conoce como error tipo II.

El primer tipo de error es el rechazo de una hipótesis nula verdadera como resultado de un procedimiento de prueba. El segundo es el fracaso en el rechazo de una hipótesis nula falsa.

Métricas obtenidas de la matriz de confusión

- Sensibilidad, recall o true positive rate (TPR): Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

- Especificidad o true negative rate (TNR): Se trata de los casos negativos que el algoritmo ha clasificado correctamente. Expresa cuán bien puede el modelo detectar esa clase.

$$TNR = \frac{TN}{TN + FP} \tag{3}$$

- Precisión, o positive predictive value (PPV): Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión, mayor la precisión. En forma práctica es el porcentaje de casos positivos detectados.

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

- Accuracy (ACC): Se refiere a lo cerca que está el resultado de una medición del valor verdadero. La exactitud es la cantidad de predicciones positivas que fueron correctas.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

- F_1-score (F_1): Esta es otra métrica muy empleada porque resume la precisión y sensibilidad en una sola métrica. Es de gran utilidad cuando la distribución de las clases es desigual. Es la media armónica entre a precisión y la sensibilidad.

$$F_1 = \frac{2TP}{2TP + FP + FN} = \frac{2PPV * TPR}{PPV + TPR} \tag{6}$$

Capítulo 3. Metodología

En este capítulo se muestra la metodología que se llevó a cabo para lograr el objetivo planteado. Se muestra geográficamente la información con la que se cuenta y se presentan los modelos predictivos propuestos, se mencionan y describen las técnicas de selección de características empleadas y los parámetros empleados para el entrenamiento de los modelos.

En el presente proyecto de tesis, se realizaron tareas de clasificación empleando el algoritmo HSP, el algoritmo con el que se comparó fue kNN. Se utilizó kNN porque es la única técnica conocida que se puede clasificar como aprendizaje basado en ejemplos.

3.1. Información general y agrupamiento de datos

La información brindada por el grupo técnico del CESVBC abarca desde junio de 2017 hasta junio de 2022. La mayoría de los datos provienen del municipio de Ensenada, y la zona con mayor información, así como mayor incidencia, corresponde a las comunidades cercanas a Valle de Guadalupe. En la Figura 16 se muestra la distribución espacial de la información con la que se dispone.

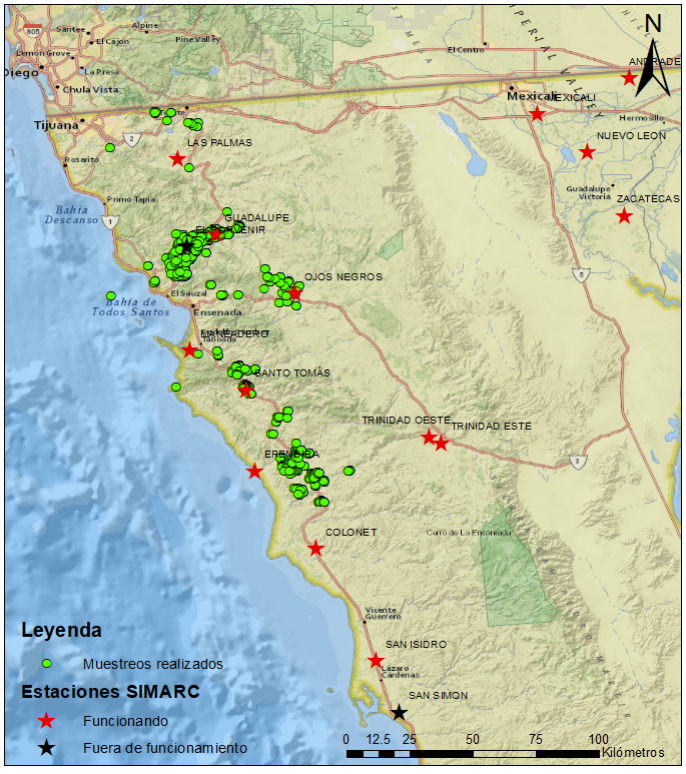

Figura 16. Distribución geográfica de los muestreos realizados en campo.

Con relación a la variedad de las vides muestreadas, la que sobresale es cabernet sauvignon (Fig. 17). En la Figura 18 se muestra como el estado fenológico de crecimiento de baya es el estado con mayor cantidad de información y con mayores niveles de infestación. Sin embargo, en todos los estados se ha encontrado la presencia del PHV (ver Fig. 55).

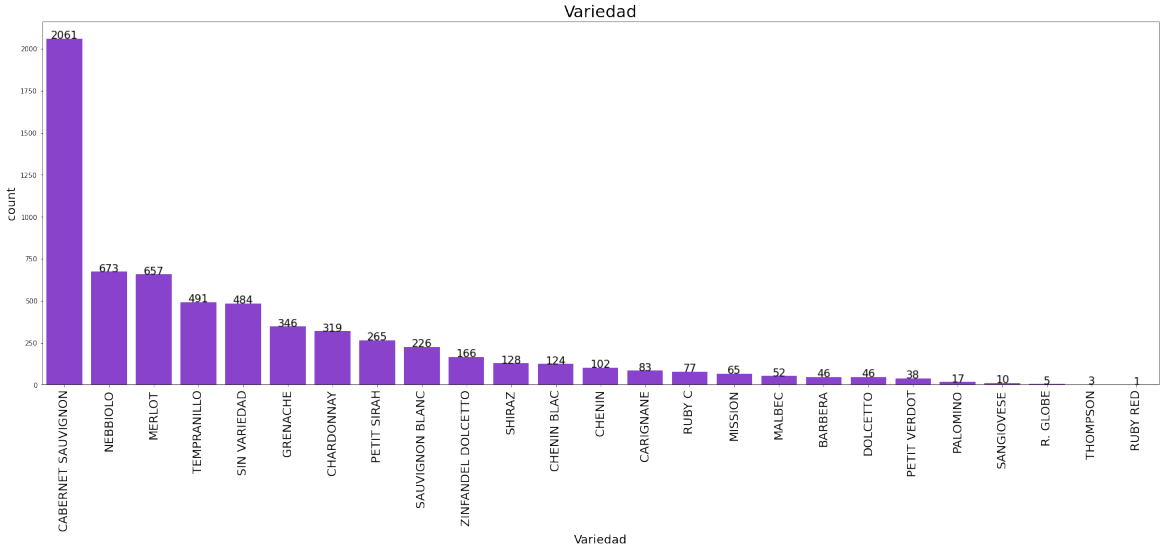

Figura 17. Variedades encontradas en los muestreos.

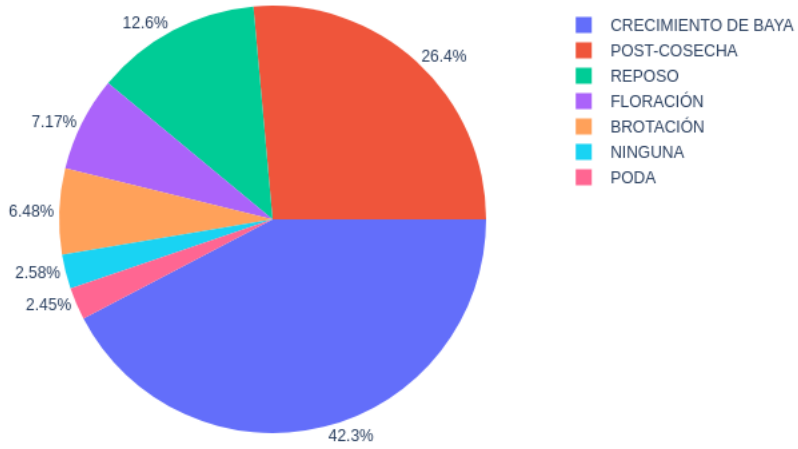

Figura 18. Fenología de las vides muestreadas.

Como se mencionó en la sección 2.2.2, y de acuerdo con la tabla 3, las variables registradas y que serán las etiquetas para el problema de clasificación son: nivel de infestación de plantas y nivel de infestación

de piojos. La primera variable se trata de la asignación que se realiza por medio del porcentaje de plantas infestadas. Esta variable tiene 4 niveles: sin presencia, nivel leve, nivel medio y nivel fuerte. En la figura 19 se aprecia la distribución de las clases, se observa que la clase sin presencia es la clase dominante. Por otro lado la variable de nivel de infestación de piojos proviene del promedio de piojos por planta, la cual también consiste de 4 niveles, pero solo se obtuvieron dos, debido a que para asignar a nivel medio y fuerte se tendría que tener un promedio mayor que 30 y 100 piojos por cada planta revisada; de hecho, independientemente del nivel asignado, en el momento que se encuentran más de 10 piojos por planta se deben tomar acciones. Con base en esto, el problema de clasificación se podría reducir a una clasificación binaria que indique la presencia o ausencia de la plaga. En este caso se observa un balance de clases, donde la clase que sobresale ligeramente del conjunto total de datos corresponde con la presencia de piojos, ver Fig. 20.

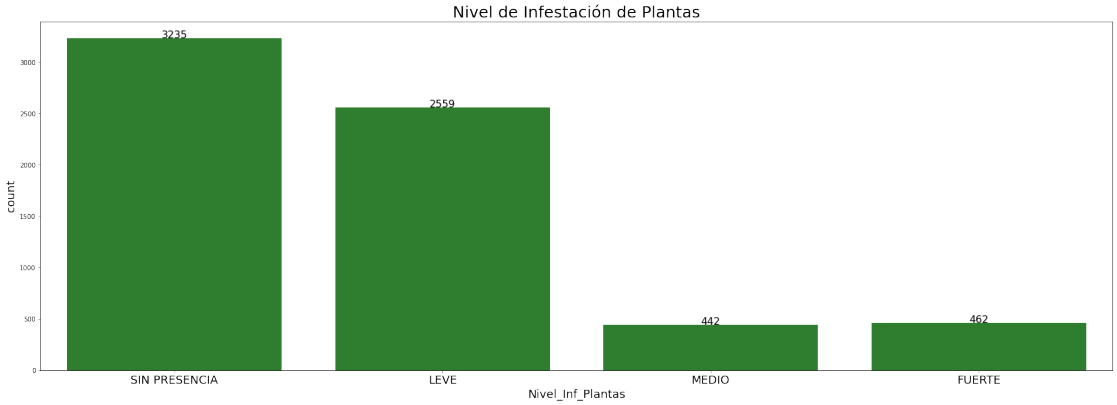

Figura 19. Distribución de clases para niveles de infestación de plantas.

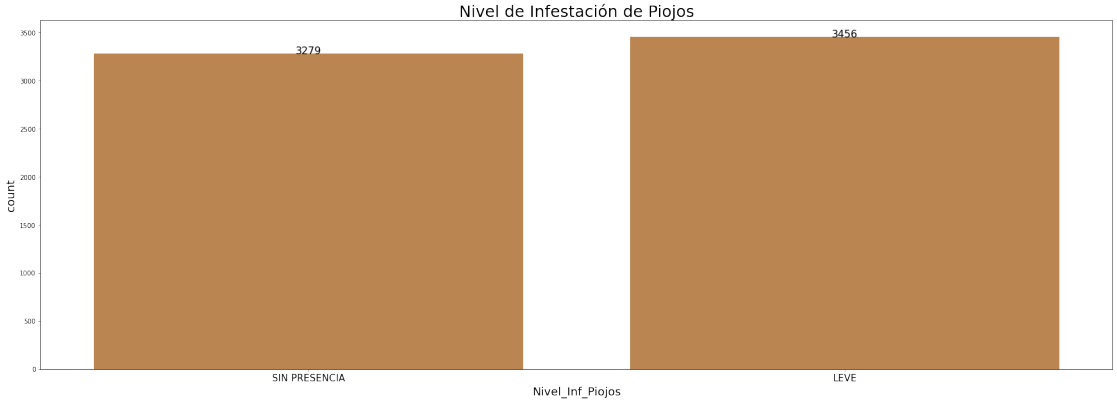

Figura 20. Distribución de clases para niveles de infestación de piojos, *i.e.*, presencia de PHV.

Con el objetivo de tener un panorama temporal, se muestra en 21 un mapa de calor donde se aprecia que en el mes de junio se han registrado las más altas tasas de incidencia, y conforme se acerca a los meses de noviembre y diciembre, disminuye la población pero no desaparece. Las celdas marcadas en

gris corresponden con falta de muestreos.

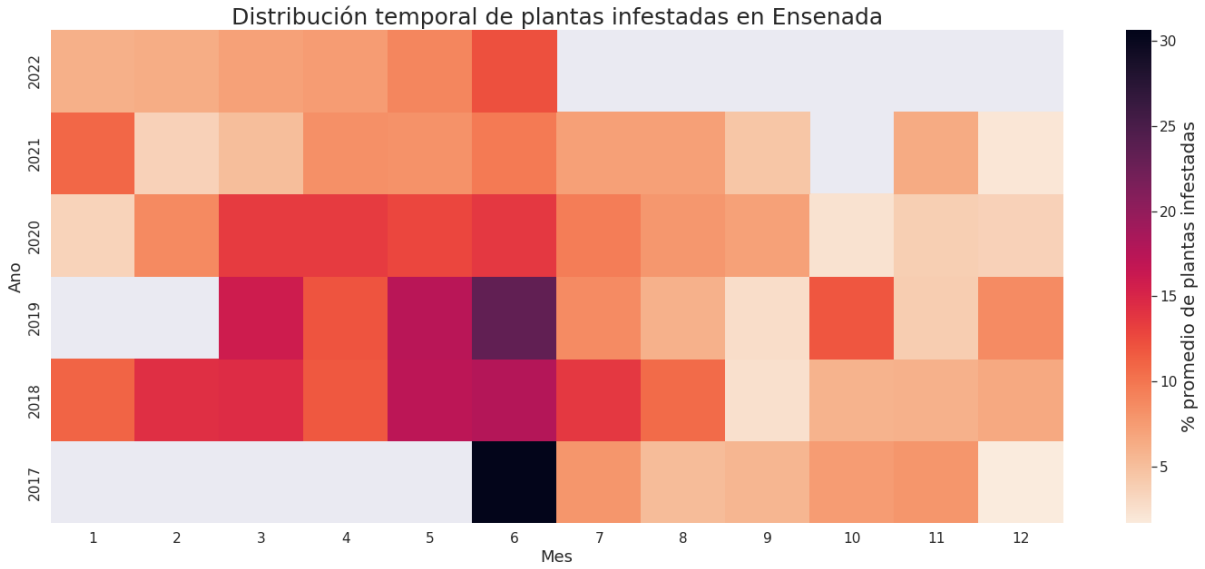

Figura 21. Distribución temporal de los muestreos realizados en Ensenada.

3.1.1. Modelos predictivos propuestos

Dado que existe gran variabilidad en la incidencia de la plaga, y considerando las diferentes condiciones de las zonas, se propuso una división de cuatro modelos predictivos, los cuales se observan en la figura 22.

- Modelo 1. Valle de Guadalupe (VG)
  - Comunidades: Francisco Zarco, Valle de Guadalupe, Valle de Calafia, Ejido el Porvenir, San Antonio de las Minas
- Modelo 2. Santo Tomás (ST)
  - Comunidades: Santo Tomás, Ejido Uruapan, Maneadero, Ejido Rodolfo Sánchez Tabodada
- Modelo 3. Ojos Negros (ON)
  - Comunidades: Ojos Negros, Real del Castillo
- Modelo 4. San Vicente (SV)
  - Comunidades: San Vicente

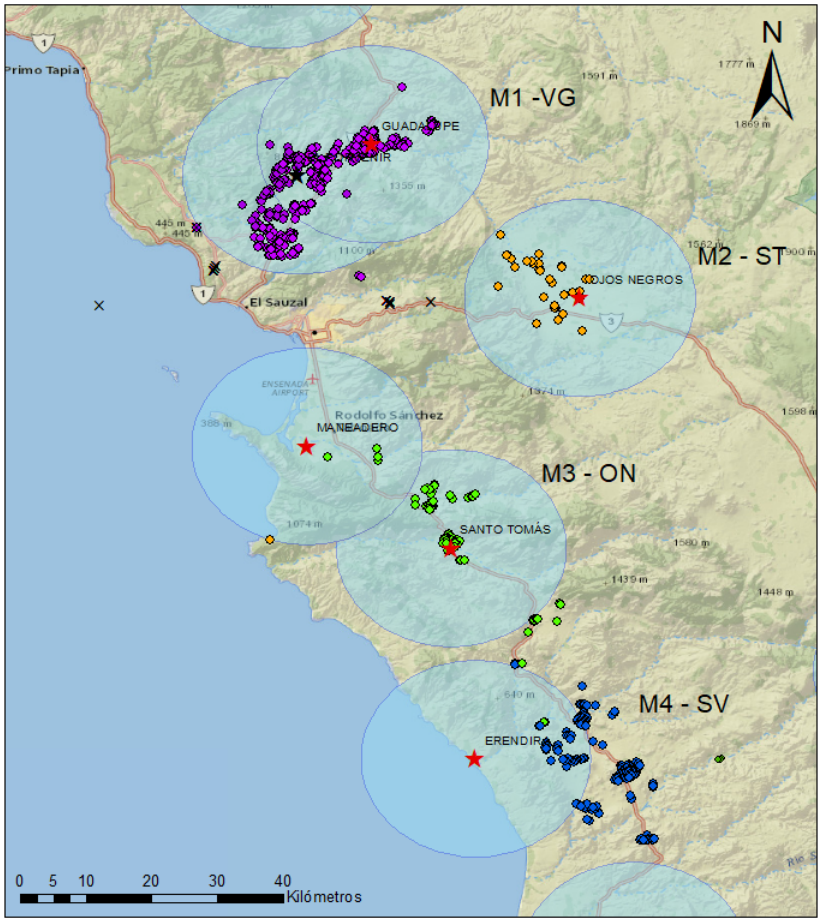

Figura 22. Modelos predictivos propuestos.

La separación de los modelos se realizó considerando un radio de 15 km a partir de las estaciones climatológicas, los datos que no entran en ese radio fueron descartados. Posteriormente, se fusionó la información de los muestreos con la información climatológica considerando la ubicación y la fecha. Cabe resaltar que el modelo 4 (San Vicente) presenta un problema: los muestreos realizados se encuentran fuera del rango de las estaciones climatológicas de libre acceso.

Dada su importancia económica, la severidad y la gran inversión que se tiene en la zona del Valle de Guadalupe, además que aquí se cuenta con la mayor cantidad de información, el proyecto y procedimiento se enfocó en el modelo 1 Valle de Guadalupe (VG). Geográficamente, el valle está localizado en el cinturón de la latitud de las regiones productoras de vino, debido a que el clima y el suelo de esta zona son precisos para el mejor aprovechamiento en el cultivo de la vid (Valladolid et al., 2018).

En las figuras 23 y 24 se observa la distribución de clases de los niveles de infestación de piojos y plantas para el modelo VG, se aprecia cómo para este modelo la clase dominante es la presencia de la plaga en

nivel leve. Por otro lado, en la figura 25 se muestra el tipo de daño, encontrado, donde se aprecia cómo la mayor parte de los piojos encontrados se localizan en el tronco, confirmando que la poca eficacia del control químico se debe principalmente a que los piojos viven debajo de la corteza y que su detección no es tarea sencilla.

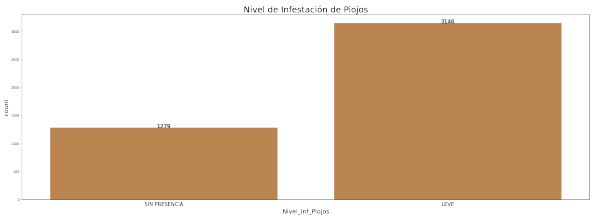

Figura 23. Niveles de infestación de PHV en VG.

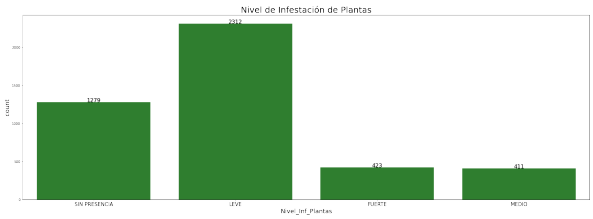

Figura 24. Niveles de infestación de plantas en VG.

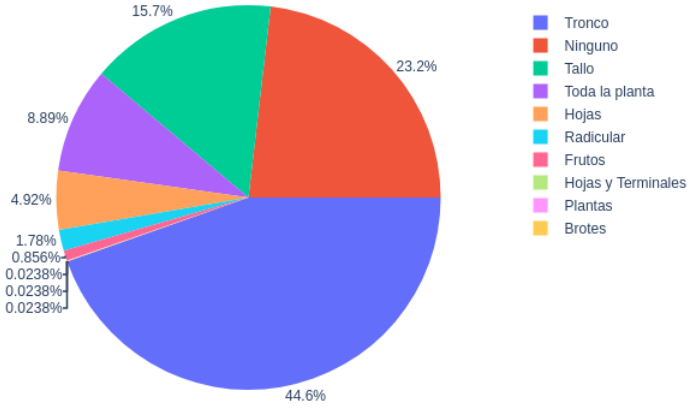

Figura 25. Tipo de daño encontrado en Valle de Guadalupe.

3.2. Selección de Características

Como se observó en la tabla 2, existe una gran variedad de características climatológicas que se emplean para realizar la predicción de alguna plaga u enfermedad. Adicionalmente, en estudios como el de Pérez-Ariza et al. (2012) se toma en cuenta características como el mes, la temporada de año, entre otras variables. Para nuestro caso de estudio, se muestran en la tabla 5 las variables climatológicas, así como características del viñedo y los datos recolectados en campo con los que se cuenta tras una previa eliminación de variables no representativas.

Tabla 5. Variables predictoras o características con las que se cuenta.

Variables climáticas	**Variables del cultivo y toma de datos**
Evapotranspiración (mm)	Fecha (año, mes, día)
Precipitación Total (mm)	Semana
Radiación Solar Total (Cal/cm^2)	Municipio y comunidad
Presión Vapor Promedio (Kpas)	Ubicación (latitud, longitud)
Temperatura Ambiente Mínima °C	Variedad
Temperatura Ambiente Máxima °C	Superficie muestreada
Media Temperatura Ambiente °C	Tipo de daño encontrado
Humedad Relativa Mínima (%)	Total de planta infestadas
Humedad Relativa Máxima (%)	Porcentaje de plantas infestadas
Media Humedad Relativa(%)	Nivel de infestación plantas
Punto de Rocío Promedio °C	Total piojos encontrados
Velocidad Media Viento (m/s)	Promedio de piojos por planta
Dirección del Viento °N	Nivel de infestación piojos
Temperatura Promedio del Suelo °C	Fenología
Unidades Calor	
Horas Frío	
Heladas	

Se aplicaron distintas técnicas de selección de características de los métodos explicados en la sección 2.3.3, las técnicas seleccionadas fueron: Selección de características basada en árboles (análisis de ganancia de información), importancia de la característica basada en la disminución media de la impureza (MDI) y el algoritmo MRMR (*Maximum Relevance Minimum Redundancy*).

La técnica basada en ganancia de información toma ventaja de los árboles de decisión, ya que este método además de ser de fácil entendimiento, calcula intrínsecamente la entropía, la ganancia de información y la impureza de Gini. La entropía (S) es un concepto que se deriva de la teoría de la información, mide la impureza de los valores de la muestra. Dado que los problemas generalmente incluyen un conjunto de características, es importante escoger bien el primer nodo del árbol de decisión debido a la pérdida de información que puede suceder; a esta pérdida se le llama entropía de la información, el objetivo entonces es buscar la manera más óptima para que la entropía sea mínima. Se define con la siguiente fórmula:

$$Entropía(S) = -\sum_{c \in C} p(c) \log_2 p(c) \tag{7}$$

donde S representa el conjunto de datos en el que se calcula la entropía, c representa las clases del conjunto S, $p(c)$ representa la proporción de puntos de datos que pertenecen a la clase c al número total de puntos de datos en el conjunto S. Los valores de entropía pueden estar entre 0 y 1. Si todas las muestras en el conjunto S pertenecen a una clase, entonces la entropía será igual a cero. Si la mitad de

las muestras se clasifican en una clase y la otra mitad en otra clase, la entropía estará en su punto más alto en 1. Para seleccionar la mejor característica y encontrar el árbol de decisión óptimo, se debe usar el atributo con la menor cantidad de entropía. La ganancia de información representa la diferencia de entropía antes y después de una división en un atributo determinado. El atributo o característica con la ganancia de información más alta producirá la mejor división, ya que clasifica los datos de entrenamiento de acuerdo con su clasificación de destino. La ganancia de información se representa con la siguiente fórmula:

$$Ganancia(S, a) = Entropía(S) - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} Entropía(S_v) \tag{8}$$

donde a representa un atributo específico o una etiqueta de clase, la entropía S es la entropía del conjunto de datos, $\frac{|S_v|}{|S|}$ representa la proporción de valores en S_V al número de valores en el conjunto de datos S. Por otro lado, la impureza de $Gini$ es la probabilidad de clasificar incorrectamente un punto de datos aleatorio en el conjunto de datos si se etiquetara en función de la distribución de clases del conjunto de datos. Dicho de otra manera, la impureza de Gini es una medida de qué tan a menudo un elemento elegido aleatoriamente del conjunto sería etiquetado incorrectamente si fue etiquetado de manera aleatoria de acuerdo a la distribución de las etiquetas en el subconjunto. La impureza de Gini se puede calcular sumando la probabilidad de cada elemento multiplicado por la probabilidad de un error en la asignación de ese elemento. Similar a la entropía, si el conjunto S es puro (es decir, pertenece a una clase), entonces su impureza es cero. Esto se denota mediante la siguiente fórmula:

$$Gini = 1 - \sum_{i} (p_i)^2 \tag{9}$$

Dependiendo del algoritmo de árbol elegido (ID3, C4.5, CART, etc) se calculan los tres, o al menos dos, de los parámetros mencionados, y además de realizar la clasificación correspondiente indica las características más relevantes, es decir, las que se encuentran más cerca de la raíz.

La técnica basada en la disminución media de la impureza también emplea árboles de decisión, pero aquí se usa un modelo de Random Forest, donde las importancias de las características se calculan como la media y la desviación estándar de la acumulación de la disminución de impurezas dentro de cada árbol. De acuerdo con Perrier (2015), MDI cuenta las veces que se usa una característica para dividir un nodo, ponderado por el número de muestras que divide; la disminución media de la impureza calcula la importancia de cada característica como la suma del número de divisiones (en todos los árboles) que incluyen la característica, proporcionalmente al número de muestras que divide.

La tercer técnica empleada fue el algoritmo de MRMR desarrollado por Ding y Peng (2003) el cual es un método mínimo-óptimo que busca identificar un pequeño conjunto de características que tengan el máximo poder predictivo posible. Por lo tanto, si una característica A y otra característica B son relevantes, pero brindan más o menos la misma información, un método de relevancia seleccionará ambas, mientras que un método mínimo óptimo seleccionará solo una de ellas y descartará la otra. MRMR funciona de forma iterativa. En cada iteración, identifica la mejor característica y la agrega a la cesta de características seleccionadas. Relevancia Máxima - Redundancia Mínima se llama así porque, en cada iteración, queremos seleccionar la característica que tiene la máxima relevancia con respecto a la variable de destino y la mínima redundancia con respecto a las características que se han seleccionado en las iteraciones anteriores (Mazzanti, 2021).

Las diez primeras características seleccionadas por cada una de las técnicas aplicadas se observan en la tabla 6.

Tabla 6. Primeras características seleccionadas con las diferentes técnicas aplicadas.

Ganancia de información	**Importancia con MDI**	**Algoritmo MRMR**
Mes	Semana	Mes
Fenología	Radiación solar total	Fenología
Temperatura del suelo	Dirección del viento	Radiación solar total
Radiación solar total	Temperatura del suelo	Temperatura media
Dirección del viento	Humedad relativa mínima	Humedad relativa media
Temperatura máxima	Temperatura mínima	Velocidad del viento
Temperatura mínima	Velocidad del viento	Precipitación total
Humedad relativa media	Fenología	Punto del rocío
Punto del rocío	Temperatura media	Año
Velocidad del viento	Humedad relativa media	Unidades calor

Tal y como se aprecia en el mapa de calor de la Figura 21 y en la Figura 26, era de esperarse que el mes fuera una característica de vital importancia, lo cual nos da un indicio sobre en qué momento se deben tomar las acciones correspondientes.

Existen variables de gran importancia, que ya se esperaban encontrar, como son la temperatura y humedad, sin embargo, otras variables como son la radiación solar total y el viento (tanto dirección como velocidad) resultaron de gran importancia. En la figura 27 se aprecia cómo al encontrarse con radiaciones mayores a $600\,Cal/cm^2$ la población del PHV empieza a incrementar. La distribución de la presencia del PHV con demás variable relevantes se pueden observar en el anexo B.

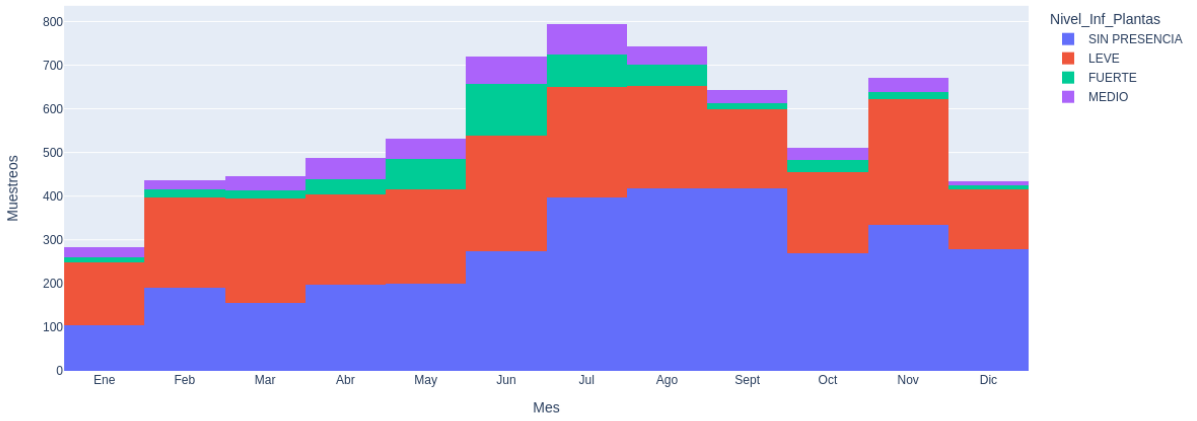

Figura 26. Distribución temporal mensual de los niveles de infestación.

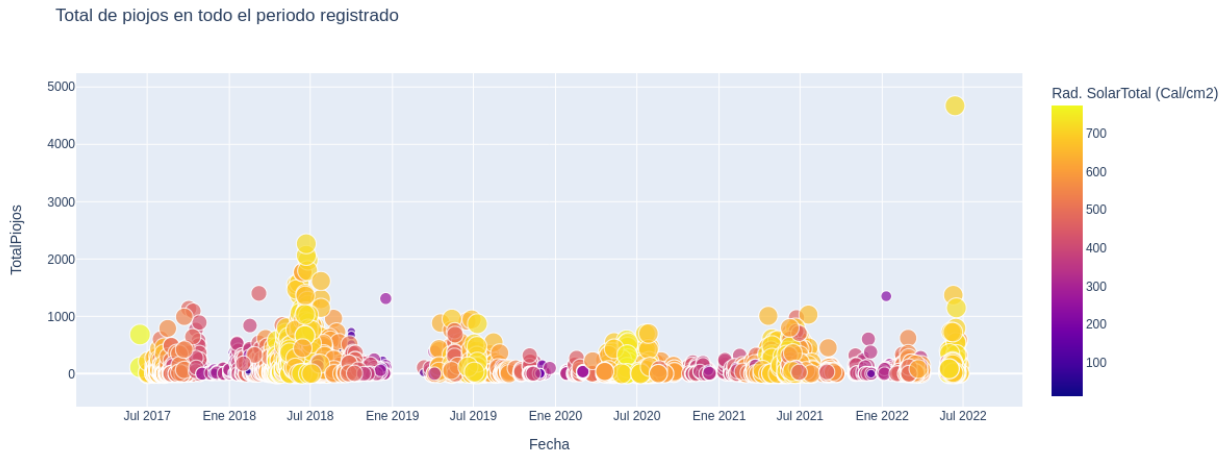

Figura 27. Número de piojos encontrados en el periodo de muestro en comparación con los niveles de radiación registrados.

3.3. Modelos de predicción

Antes de realizar los entrenamientos y sus evaluaciones, los conjuntos de datos de los modelos se prepararon reduciendo las características a una escala normalizada. La normalización consiste en ajustar los valores medidos en diferentes escalas a una escala común. Si los datos no se normalizan, se tienen diferentes escalas, lo que puede causar que algunas características sean dominantes sobre otras características (Trebuna et al., 2014), por ejemplo, la variable de precipitación está medida en milímetros, mientras que las variables de temperatura están en grados centígrados. El objetivo de la normalización es asignar a cada característica un valor dentro de un rango de 0 a 1. De igual manera, a las variables categóricas (*e.g.*, la fenología) se les debe aplicar una transformación; esto se conoce como codificación

categórica, el cual es un proceso de conversión de categorías a números. La técnica que aplicó fue la codificación *One Hot*; donde crean características adicionales basadas en el número de valores únicos de esa categoría, y cada categoría se representa como un vector conocido como *one-hot*, tal y como se muestra en tabla 7.

Tabla 7. Ejemplo de codificación one-hot.

Color
Rojo
Verde
Azul
Verde

Característica

Rojo	Verde	Azul
1	0	0
0	1	0
0	0	1
0	1	0

Codificación one-hot

3.3.1. Clasificación con kNN

Dos modelos fueron entrenados para clasificar el nivel de infestación de plantas (4 clases) y la presencia de piojos (2 clases). Dada la naturaleza del clasificador kNN, se tuvo que entrenar previamente n veces ($n = 25$) cambiando el número de vecinos $k = 1, 2...n$, para elegir el valor de k, ver Figuras 29 y 30.
El clasificador kNN puede ser susceptible a valores atípicos, mejor conocidos como outliers, en especial si el valor elegido de k es pequeño o cercano a uno; por otro lado, si el valor de k es alto se puede decir que el modelo es robusto ante outliers. Por ejemplo; con $k = 1$, supongamos que hay 1 outlier cerca de nuestro punto de prueba y el modelo predice la etiqueta correspondiente a ese outlier. En este mismo escenario, si tomamos $k = 7$, hay otros 6 vecinos más cercanos (no outliers) y 1 outlier en la vecindad del punto de prueba, y cuando se toma el voto mayoritario, se obtiene el resultado basado en los 6 vecinos más cercanos. Para lidiar con la presencia de los outliers se aplicó el algoritmo Local Outlier Factor (LOF), propuesto por Breunig et al. (2000), el cual es un método de detección de anomalías no supervisado que calcula la desviación de densidad local de un punto de datos dado con respecto a sus vecinos. Considera como valores atípicos las muestras que tienen una densidad sustancialmente menor que sus vecinos.

Otro paso necesario previo al entrenamiento corresponde con tratar el desbalance de clases, como se observa en las Figuras 23 y 24, donde existe un desbalance considerable, este escenario es muy común en distintas tareas, por ejemplo, en la detección de un tumor, detección de spam, etc. Ante estos escenarios se pueden aplicar diversas técnicas: colectar más información, cambiar las métricas de rendimiento, hacer re-muestreo del conjunto de datos, generar muestras sintéticas, probar algoritmos diferentes, probar modelos penalizados, o bien, probar una perspectiva diferente. En el presente proyecto se optó por el re-

muestreo aleatorio del conjunto de datos de entrenamiento. Los dos enfoques principales para muestrear aleatoriamente un conjunto de datos desbalanceados son eliminar ejemplos de la clase mayoritaria, lo que se denomina submuestreo, y duplicar ejemplos de la clase minoritaria, lo que se denomina sobremuestreo. Para los entrenamientos realizados, se empleó una técnica de sobremuestreo aleatorio para incrementar la cantidad de ejemplos de las clases minoritarias.

En las figuras 29 y 30, se observa el rendimiento del modelo de Valle de Guadalupe entrenado para diferentes valores de k. Se señala con una estrella azul el valor de *accuracy* más alto para la etapa de prueba. Si consideramos esto, el valor elegido sería $k = 3$ para ambas clasificaciones, por otro lado, si se toma en cuenta el sobreajuste, se denomina sobreajuste u *overfitting* al hecho de hacer un modelo tan ajustado a los datos de entrenamiento que haga que no generalice bien a los datos de prueba (ver Fig. 28), por ejemplo, se podría elegir un valor más alto *e.g.* $k = 13$. En el siguiente capítulo se ahonda más sobre la elección del valor de k, ya que debido al tipo de problema, nos interesa reducir los falsos negativos (ver Figura 15). Un falso negativo corresponde con una predicción negativa cuando lo real es positivo, *i.e.*, se predijo que no se tiene incidencia de PHV cuando la realidad es que sí hay. En este sentido, se tiene una pérdida directa pues no se tomaron las acciones correspondientes. En contraste, los falsos positivos corresponden a predicciones positivas de la presencia de plaga cuando la realidad es que no hay, en este caso se pueden tomar medidas preventivas las cuales no resultan perjudiciales para el cultivo.

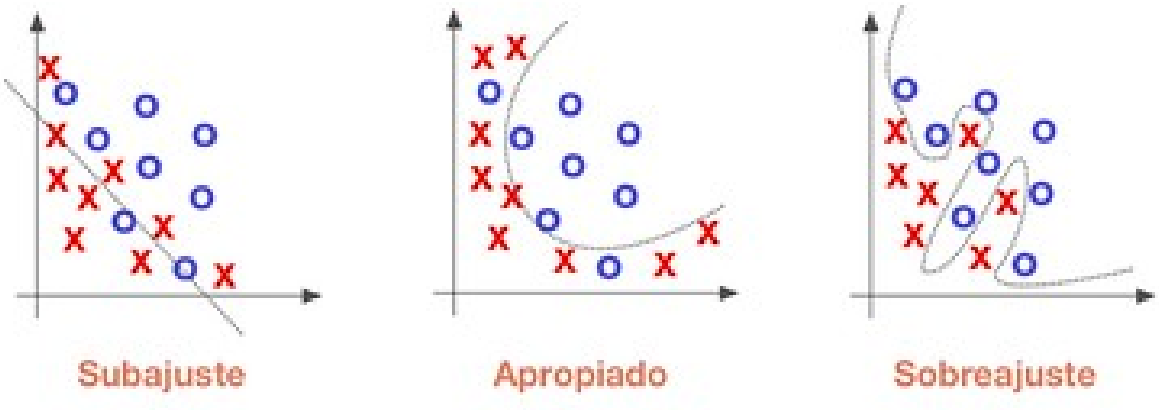

Figura 28. Representación de subajuste, ajuste apropiado y sobreajuste para una clasificación binaria.

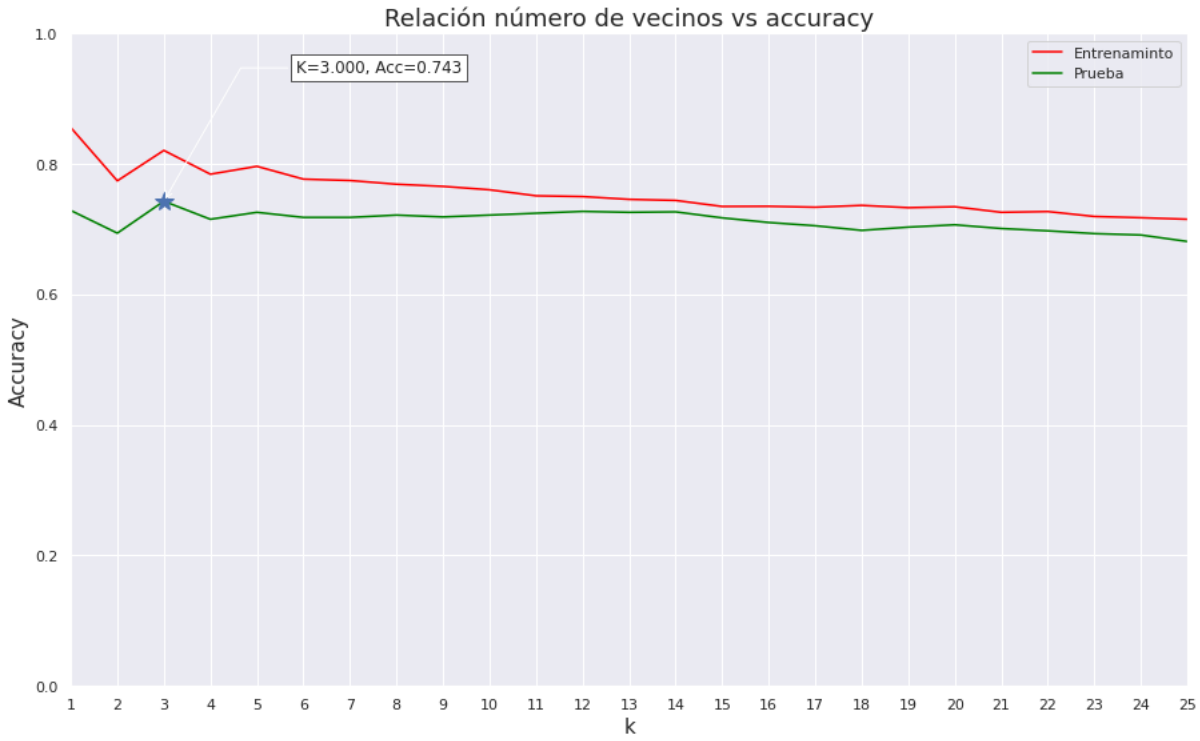

Figura 29. Accuracy del clasificador kNN con diferentes valores de K, presencia de piojos.

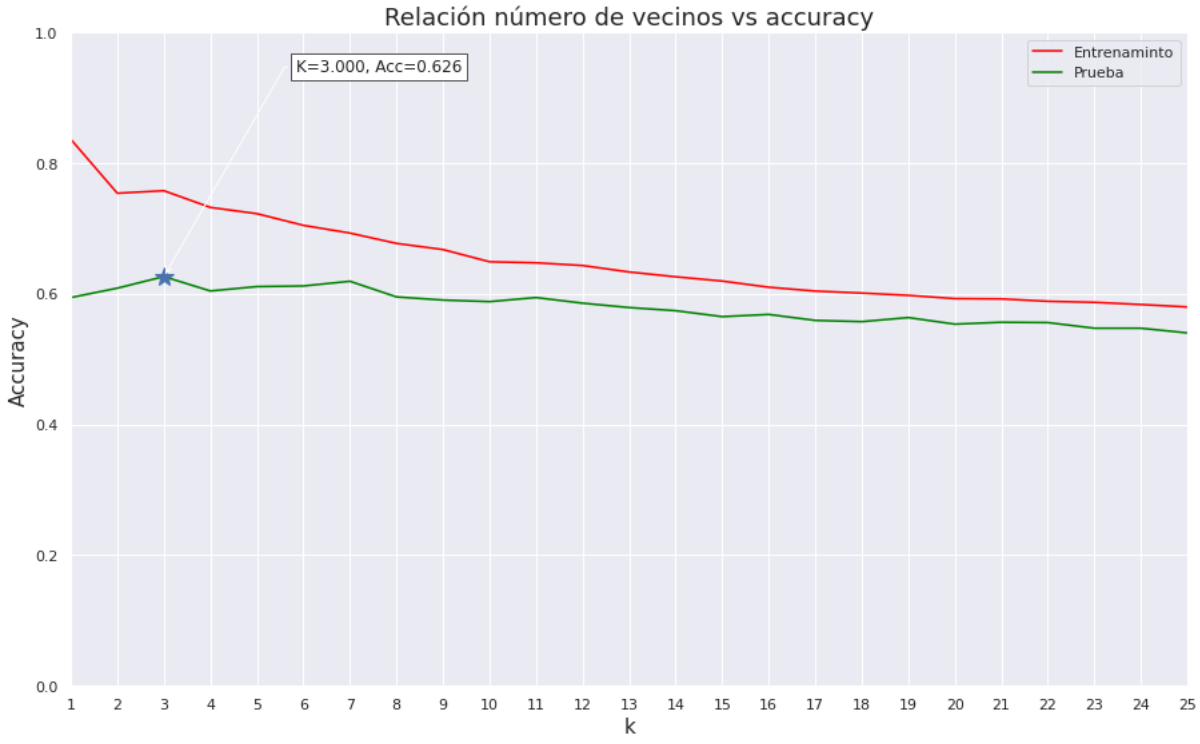

Figura 30. Accuracy del clasificador kNN con diferentes valores de K, nivel de infestación.

3.3.2. Clasificación con HSP

Las mismas técnicas fueron aplicadas para el entrenamiento empleando el algoritmo HSP, *i.e.*, se quitaron outliers con el algoritmo LOF y se aplico sobremuestro aleatorio a las clases minoritarias. La principal y gran ventaja de estos modelos es que no se debe entrenar n veces para elegir el número de ejemplos (vecinos). Tanto el entrenamiento, o mejor dicho, la clasificación del conjunto de prueba, se realiza una única vez sin la necesidad de elegir algún parámetro previo.

El número de ejemplos seleccionados para asignar la clasificación varia en cada caso: hay instancias que sólo tienen un ejemplo cerca por lo que se asigna a esa clase, mientras que hay otros datos de prueba que el algoritmo seleccionó hasta 10 ejemplos. Con estas instancias seleccionadas, que no necesariamente son las más cercanas dada la selección de los nodos (ver Figura 11), se les aplica una regla de votación simple que consiste en elegir al elemento más votado; para los casos donde existe un empate entre clases, el desempate consistió en seleccionar los primeros ejemplos impares, *i.e.*, 1,3,5,7, según sea el caso, con eso se le asigna más importancia a los primeros elementos, al mismo tiempo que se evita el empate.

En el siguiente capítulo se muestran los resultados obtenidos de los entrenamientos realizados con los clasificadores kNN y HSP para el modelo de Valle de Guadalupe, tras aplicar la metodología descrita.

Capítulo 4. Resultados

El objetivo principal de esta tesis es desarrollar una metodología de apoyo a la toma de decisiones que facilite la predicción sobre posibles apariciones de plagas o enfermedades en algún cultivo monitorizado empleando técnicas de aprendizaje de máquina.

La plaga que se estudió es el piojo harinoso de la vid, una plaga insectil de reciente aparición en territorio nacional, que representa una grave amenaza para la viticultura. Es uno de los insectos más dañinos y difíciles de controlar en el complejo tradicional de plagas que atacan a la vid. Presenta grandes pérdidas en regiones vitivinícolas de Ensenada, donde el área con mayor preocupación corresponde con el Valle de Guadalupe. Los modelos de aprendizaje empleados, cuyos resultados se muestran en el presente capítulo, son kNN y HSP. Ambos modelos se basan en ejemplos, son de fácil implementación y explicables; la principal motivación de emplear este tipo de aprendizaje es que utiliza un modelo implícito, donde se almacena la información y la predicción se realiza hasta el momento de una nueva consulta.

Es la primera vez que se utiliza aprendizaje de máquina para analizar la información recolectada en Ensenada, los resultados aquí expuestos se pueden considerar como punto de partida para cualquier estudio posterior. La metodología desarrollada, si bien sigue el procesamiento clásico de análisis de datos, se presenta como una oportunidad para procesar la información de muestreos recolectados en campo siguiendo las metodologías establecidas por los comités estatales de sanidad del país.

La recolección de datos aquí reportados inició en junio de 2017. La información contenida es incompleta debido a diversos factores cuyo análisis escapa al alcance de esta tesis. En la Figura 31 se muestra la cantidad de información recolectada anualmente, se aprecia cómo en los años 2019 y 2020 se tiene un déficit de información, lo cual sucedió por falta de personal en campo debido a la pandemia del Covid-19, así como por insuficiencia presupuestal de acuerdo con el CESVBC. Los años 2018 y 2021 cuentan con información completa, es decir, todos los predios analizados cuentan con cuatro muestreos anuales. Otro tipo de problemas corresponden al escaso número de estaciones climatológicas, el no funcionamiento de las existentes, así como la inaccesibilidad a información de ciertas redes.

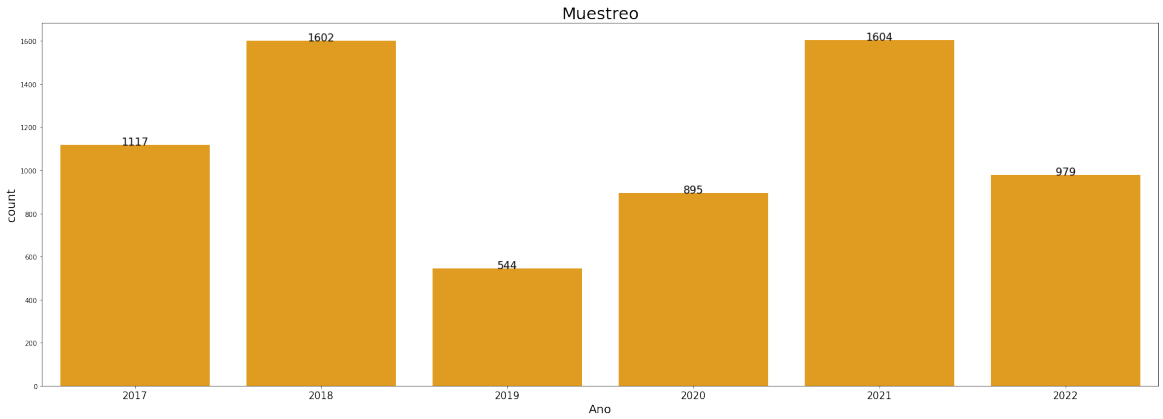

Figura 31. Cantidad de muestreos realizados por año.

Como se explicó en la sección 2.3.4, una manera de visualizar el desempeño de los algoritmos es la matriz de confusión, recordando que para la tarea de clasificación la diagonal principal muestra las clasificaciones correctas, y fuera de la diagonal, las confusiones correspondientes con falsos positivos y falsos negativos. Es importante resaltar que buscamos que se disminuyan los falsos negativos, ya que con esta predicción errónea, predecir que no hay plaga cuando sí la hay, provocaría una pérdida y afectación directa ante la falta de acciones. Por otro lado, un falso positivo, *i.e.*, no se tiene la presencia de plaga pero se predijo que sí, se pueden tomar acciones las cuales implican un gasto económico, lo cual sería preferible que la pérdida de cultivo.

4.1. Resultados con kNN

El algoritmo de k vecinos cercanos es un clasificador dentro del tipo de aprendizaje supervisado, en específico basado en ejemplos o instancias, que utiliza la proximidad para hacer clasificaciones partiendo de la suposición de que si se tiene una muestra cuya etiqueta no se conoce, se puede predecir la etiqueta por las etiquetas que hay dentro de una vecindad de la consulta, la vecindad está dada por los k vecinos cercanos.

Como se ha mencionado, el principal problema del algoritmo kNN es la correcta elección del valor de k. A continuación se muestran las matrices de confusión obtenidas con diferentes valores de k para clasificar la presencia de piojos en el Valle de Guadalupe.

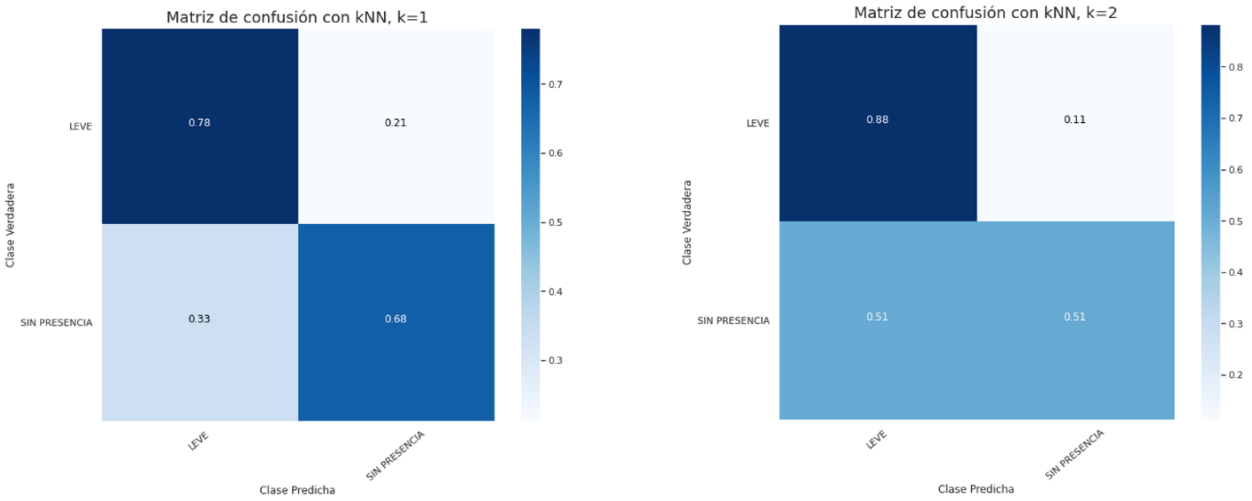

Figura 32. Matrices de confusión normalizadas con kNN para presencia de piojos con k=1, 2

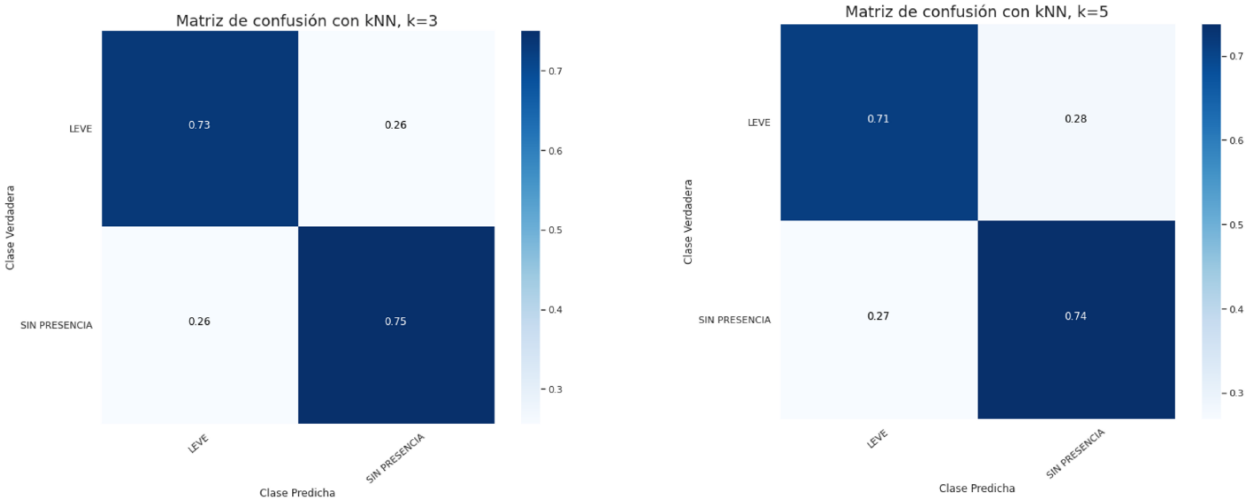

Figura 33. Matrices de confusión normalizadas con kNN para presencia de piojos con k=3, 5

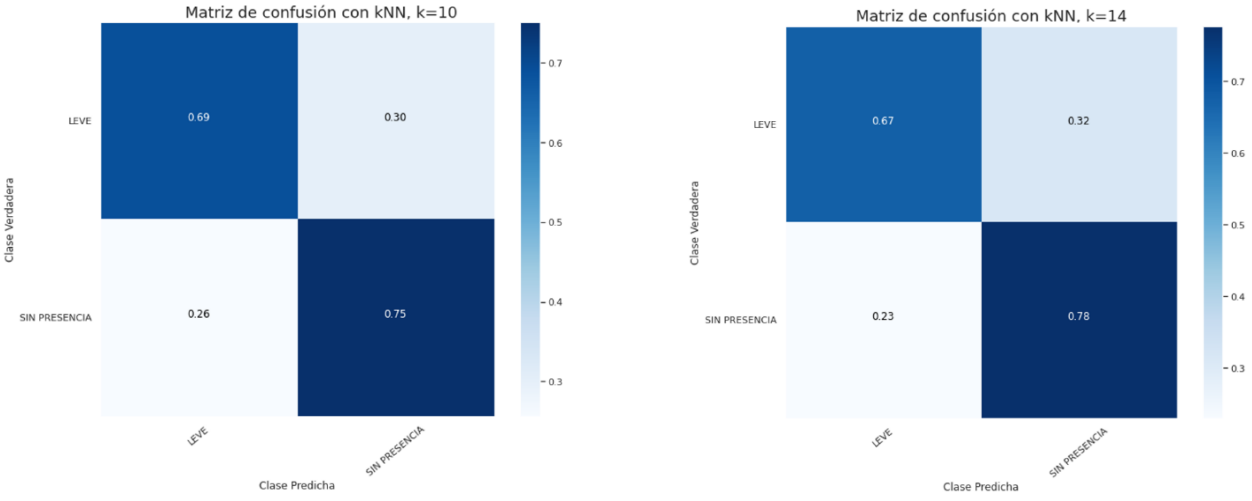

Figura 34. Matrices de confusión normalizadas con kNN para presencia de piojos con k=10, 14

Se observa en las matrices, y gracias a la escala de color, cómo las predicciones realizadas con un valor de $k = 2$, se obtiene la menor cantidad de falsos negativos, aunque el accuracy de este valor está por debajo de otros (acc=0.67) y se tiene una gran cantidad de falsos positivos. Por otro lado, con valores de $k = 1$ ó 3, el accuracy aumenta alcanzando valores de $74\,\%$ de correctas clasificaciones, pero se tiene un ligero incremento de falsos negativos. En las Figuras 32 - 34 se aprecia cómo al aumentar el valor de k, *i.e.*, $k = 5, 10, 14$, se tienen la mayor cantidad de falsos negativos, lo cual es precisamente lo que se desea evitar, por lo que se descarta la posibilidad de elegir valores de k mayores a 5, aún considerando que el sobreajuste disminuye.

En la Figura 35 se muestran las métricas obtenidas con $k = 1$. Se observa que el valor de sensibilidad para la clase que indica la presencia de la plaga es de 0.78, mientras que con $k = 2$ (fig 36) se tiene el mayor valor de sensibilidad (0.88). Por otro lado, para $k = 3$ se tiene el accuracy más alto (acc= 0.74) y la sensibilidad de clases se estabiliza con un $73\,\%$ a $75\,\%$ de clasificaciones correctas para ambas clases (fig 37).

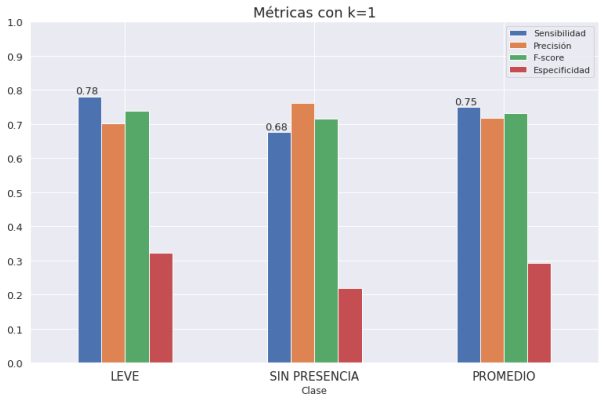

Figura 35. Métricas con kNN para presencia de piojos, k=1.

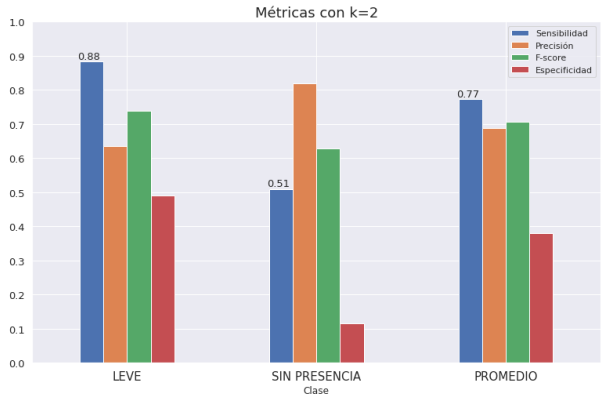

Figura 36. Métricas con kNN para presencia de piojos, k=2.

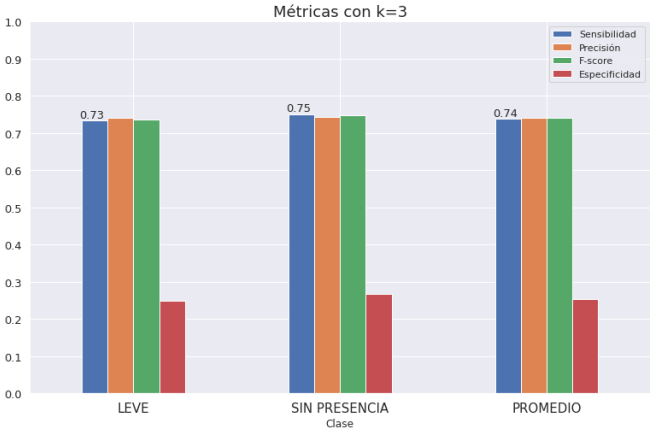

Figura 37. Métricas con kNN para presencia de piojos, k=3.

La segunda tarea de clasificación determina el nivel de infestación: consiste de una clasificación de cuatro niveles desde la clase sin presencia, leve, medio hasta infestación fuerte. A continuación se muestran las matrices de confusión obtenidas para distintos valores de k.

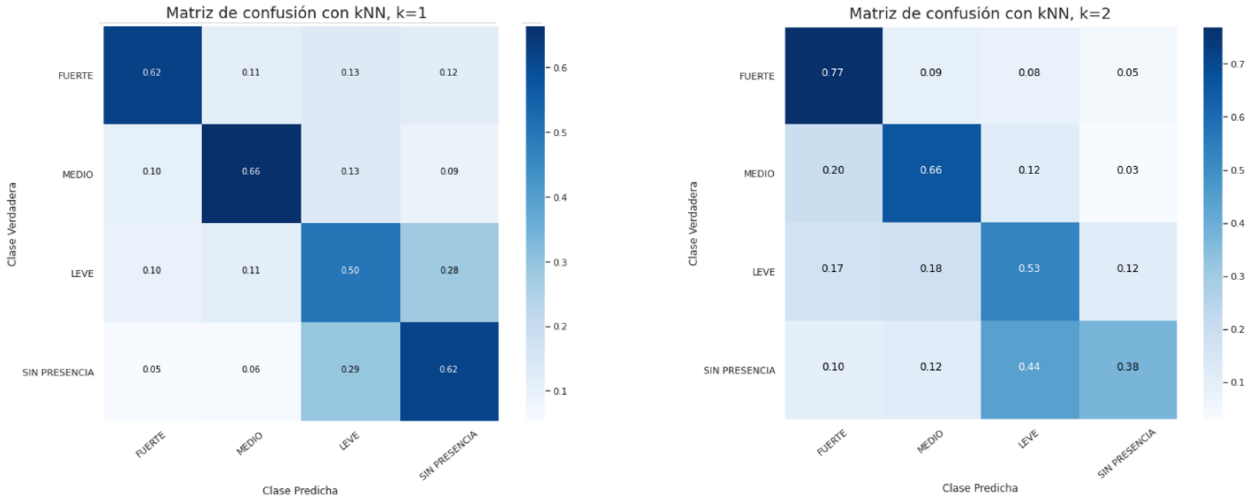

Figura 38. Matrices de confusión normalizadas con kNN para niveles de infestación con k=1, 2.

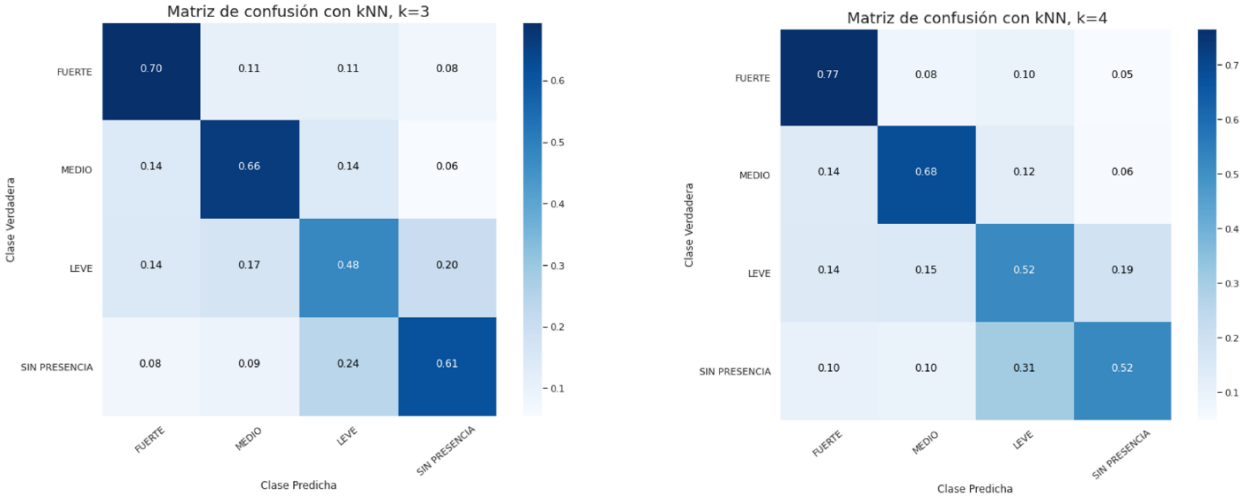

Figura 39. Matrices de confusión normalizadas con kNN para niveles de infestación con k=3, 4.

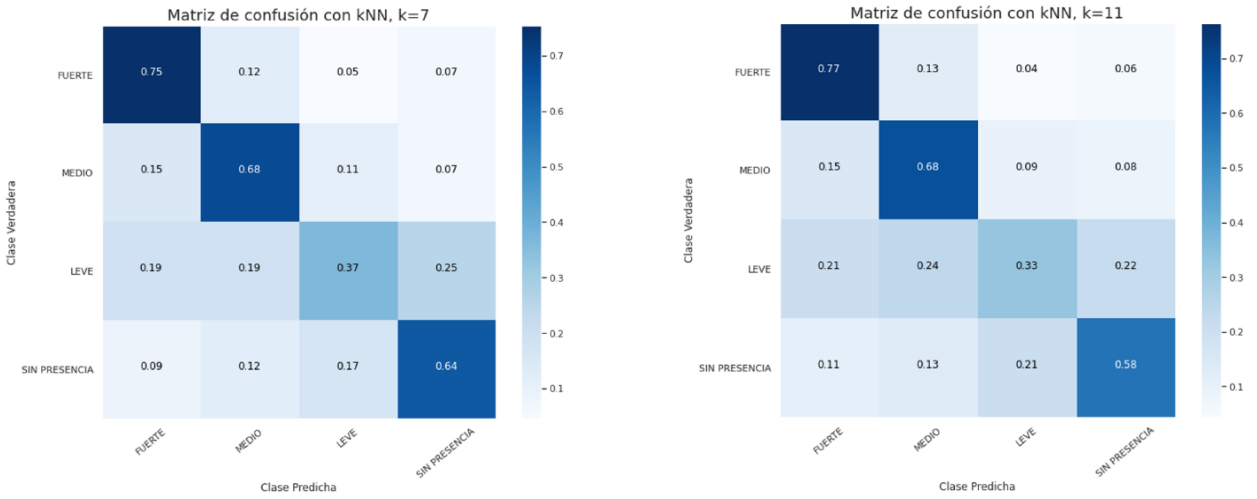

Figura 40. Matrices de confusión normalizadas con kNN para niveles de infestación con k=7, 11.

Como ahora se trata de cuatro posibles resultados, se tiene mayor variabilidad entre clases y se aumentan las confusiones, sin embargo, las clases extremas, *i.e.*, el nivel fuerte y sin presencia son las clases que más se diferencian y entre ellas existen menos confusiones, mientras que las clases que más confusiones presentan son aquellas sin presencia, con el nivel leve. Naturalmente se tiene la presencia de falsos positivos y negativos, de igual manera que con la clasificación binaria, con un valor de $k = 2$ se tienen la menor cantidad de falsos negativos, independientemente de que el accuracy con este valor no es el más alto, el accuracy más alto ocurre con $k = 3$ (ver Fig. 30).

En las Figuras 41 y 42 se presentan sus correspondientes métricas para las clasificaciones realizadas con k= 2 y 3. Se observa que para la clase fuerte se tienen sensibilidades de 0.77 y 0.70, respectivamente.

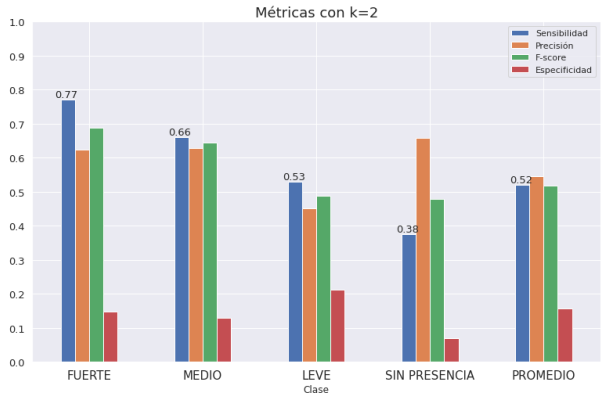

Figura 41. Métricas con kNN para nivel de infestación, k=2.

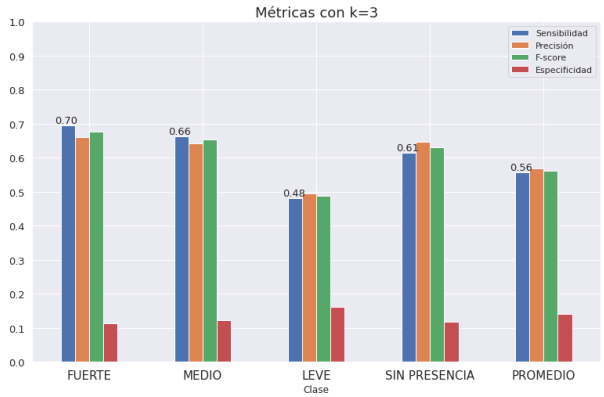

Figura 42. Métricas con kNN para nivel de infestación, k=3.

El mayor valor de sensibilidad corresponde con la clase fuerte, lo cual nos beneficia, ya que se tiene mayor

seguridad de predicción para la presencia de la plaga. Las clases que más se confunden son el nivel leve con la clase sin presencia, lo cual puede ser indicio de la ligera transición que se tiene al inicio de la plaga.

4.2. Resultados con HSP

El algoritmo HSP (Talamantes y Chavez, 2022) también se trata de un algoritmo basado en ejemplos el cual conserva la simplicidad del kNN con la ventaja de ser libre de parámetros, es decir, no se tiene que definir el valor de k. En esta sección se muestran los resultados obtenidos de los modelos entrenados con el clasificador HSP.

En la Figura 43 se muestra la matriz de confusión obtenida para clasificar la presencia del PHV en el Valle de Guadalupe, y en la Figura 44 se muestran las métricas obtenidas. Se observa que los valores de sensibilidad son similares a los obtenidos con KNN: $73\,\%$ de correctas clasificaciones; de manera general se aprecia un buen desempeño del clasificador, aún considerando la omisión de los entrenamientos previos y con la ventaja de ser un modelo libre de parámetros.

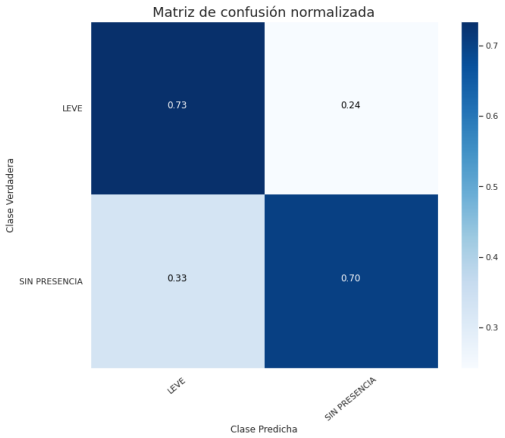

Figura 43. Matriz de confusión con HSP para presencia de piojos.

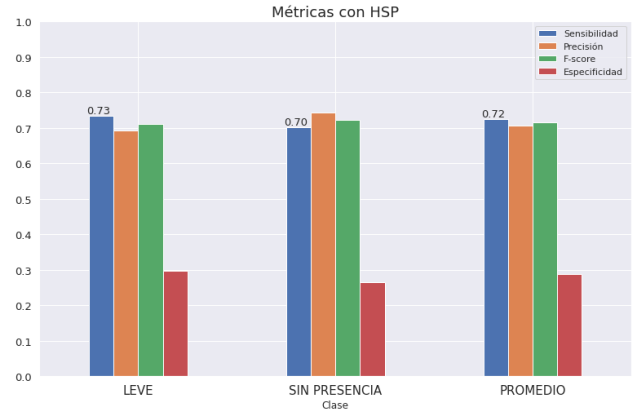

Figura 44. Métricas con HSP para presencia de piojos en modelo 1 VG.

En las Figuras 45 y 46 se muestra la matriz de confusión y las métricas obtenidas para la clasificación de los niveles de infestación. Nuevamente, las clases que más se diferencian son los extremos (clase fuerte y sin presencia), se tienen más confusiones en la transición al inicio de la plaga y se tiene un balance entre la presencia de falsos positivos y negativos.

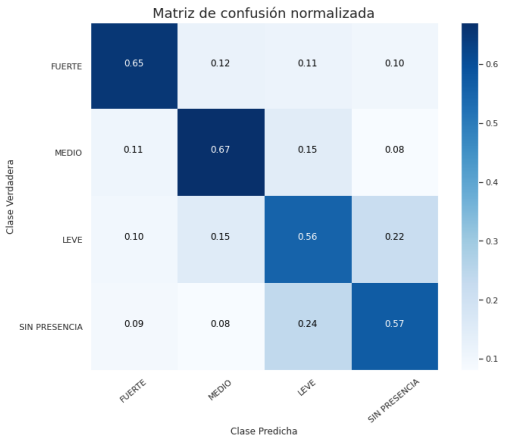

Figura 45. Matriz de confusión con HSP para niveles de infestación.

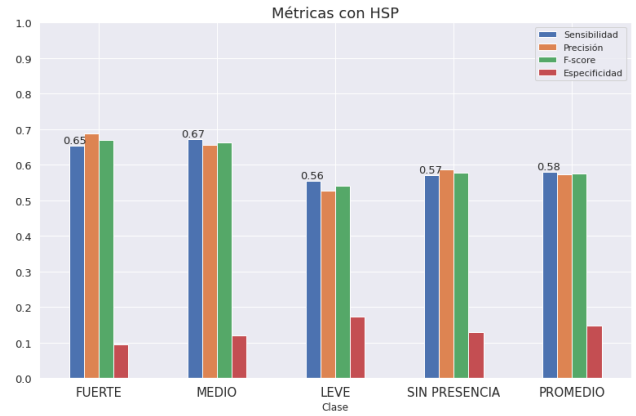

Figura 46. Métricas con HSP para niveles de infestación en modelo 1 VG.

4.3. Discusión de resultados

La predicción de la presencia del PHV se realizó empleando dos algoritmos de clasificación basados en ejemplos: kNN y HSP. De manera general ambos algoritmos obtuvieron métricas similares, sin embargo, se debe considerar la incertidumbre que se tiene en la elección del valor k y que se debe tomar una heurística para determinar su valor, *e.g.*, cuando los datos están etiquetados es fácil determinar el valor de k, pues se puede realizar una comparación, sin embargo, cuando se tiene información nueva no etiquetada no hay manera de saber cuál es la k adecuada.

Los valores de k que obtuvieron métricas más elevadas y menos falsos negativos fueron valores pequeños ($k = 1, 2, 3, 4$), sin embargo, se observa en las Figuras 29 y 30 que en estos valores es donde se tiene mayor sobreajuste, lo cual es indicio que cuando se tenga nueva información, por ejemplo, la información recolectada en el segundo semestre de 2022, el modelo no podrá generalizar bien pues se sobrejustó a la información ya vista. Por otro lado, con el modelo HSP, cuando se decida probar con información nueva se tiene la ventaja de no tener que decidir nada a priori, pues la información estará almacenada para cualquier consulta posterior.

Considerando la información disponible para los modelos presentados y con el objetivo de robustecer los modelos en el futuro, se recomienda que la información siga siendo colectada bajo la misma metodología, con más años de registro, bajo las mismas condiciones posibles y evitando problemas externos. Por ejemplo, Chen et al. (2020) emplearon información recolectada durante 9 años para predecir la incidencia de Cenicilla vellosa en viñedos de Francia. Otro ejemplo es el estudio de Pérez-Ariza et al. (2012), quienes además de emplear información recolectada por 8 años, se trata de una granja experimental para analizar

la roya del café.

En relación con los problemas externos se puede hablar, por ejemplo, de la poca cobertura así como de la falta de supervisión y reparación oportuna de las estaciones agroclimáticas; esto es un problema externo pero que tiene un impacto negativo para estudios de este tipo, pues hay diversos meses donde sí se cuentan con muestreos pero no se tiene información de las condiciones climáticas, o bien, hay regiones como lo es la comunidad de San Vicente donde hay presencia e incremento de la plaga (ver Figuras 22 y 50) pero no se cuenta con ninguna estación cercana de libre acceso.

Una situación relacionada con la metodología fue la mencionada falta de personal en ciertos periodos de tiempo, lo cual resultó en un déficit considerable de muestreos. Afortunadamente, de acuerdo con el CESVBC, la recolección de más información, *i.e.*, los muestreos, se sigue llevando a cabo y la información se almacena y actualiza constantemente en las oficinas y servidores pertinentes.

El problema abordado en el presente proyecto de tesis tiene como finalidad principal el brindar recomendaciones basadas en la información analizada, es decir, si bien las métricas son un buen indicador del desempeño de los modelos, lo que se busca es transmitir el conocimiento adquirido de alguna manera en que se puedan llevar a cabo ciertas acciones. Ante este escenario se plantean recomendaciones mencionando previamente las posibles pérdidas de cultivo que se pueden tener de acuerdo al nivel de infestación.

De acuerdo con el Anuario Estadístico de la Producción Agrícola (SIAP, 2021), en el año 2021 el estado de Baja California reportó una producción de $24,748.05$ toneladas de uva industrial en un superficie cosechada de $3,864$ hectáreas. El rendimiento por hectárea fue de 6.40 toneladas con un costo de producción de $\$16,040.70$.

Recordando que los muestreos se realizan en lotes de 10 hectáreas y con la información mencionada, se puede suponer que un lote tiene un rendimiento óptimo de 64 toneladas aproximadamente. Considerando esto, y los niveles de infestación descritos en la tabla 3, se muestra en la tabla 8 tres escenarios posibles (mínimo, promedio y máximo) de la cantidad de posibles toneladas infestadas de acuerdo con nivel de infestación reportado por lote.

Tabla 8. Toneladas de uva infestadas por lote de acuerdo al nivel de infestación considerando escenario mínimo, promedio y máximo.

Nivel de infestación	Escenario	% de infestación	Toneladas infestadas
Sin presencia	-	0 %	0
Leve	Mín	1 %	0.64
	Prom	11 %	6.72
	Máx	20 %	12.8
Medio	Mín	21 %	13.44
	Prom	26 %	16.32
	Máx	30 %	19.2
Fuerte	Mín	31 %	19.84
	Prom	66 %	41.92
	Máx	100 %	64

4.4. Recomendaciones

- Una de las principales sugerencias que se tienen ante la falta de datos, o cuando se tienen clases desbalanceadas, es saber si se pueden colectar más información. Afortunadamente en el caso de este estudio, se sigue recolectando información, la cual podrá ser empleada en el futuro para robustecer los modelos aquí mostrados. Por lo que seguir con la recolección de datos y en especial mantener el vínculo con el comité, será clave para seguir estudiando la dinámica de la plaga y su gestión.
- En la medida de lo posible, estar al pendiente del correcto funcionamiento de las estaciones agroclimáticas, así como realizar una petición para reparar las existentes: la estación El Porvenir (ver Fig. 7) lleva sin funcionar desde el año 2015, y por su ubicación habría sido muy enriquecedor poder trabajar con esa información. Por otro lado, existen redes de estaciones privadas, en este proyecto se realizó una petición, sin embargo, no resultó satisfactorio ya que solo compartieron ciertos años de registro, y para estudios como el presente se requiere de una red a la que constantemente se pueda acceder a la información.
- Con relación a los modelos de aprendizaje de máquina se recomienda continuar con técnicas basadas en ejemplos, ya que este tipo de técnicas requieren un bajo costo computacional, son de fácil entendimiento y se minimiza la necesitad de supervisión de un científico de datos.

A continuación se describen otro tipo de recomendaciones relacionadas con el manejo de la plaga:

- A pesar que se sabe que se tienen generaciones superpuestas, se observó que la mayor población de PHV en el Valle de Guadalupe ocurre en los meses de junio y julio en todos los años de registro,

y que comienza a disminuir en el mes de septiembre. Ante esto se recomienda realizar acciones en los meses previos a los mayores picos de infestación (ver Figs. 26 y27).

- De acuerdo con las técnicas de selección de características implementadas: el mes, la fenología, la radiación solar total y la temperatura del suelo, son las características más importantes a la hora de realizar las predicciones; en particular la radiación solar demostró ser un indicador de gran importancia. Al alcanzar radiaciones mayores a $600\ cal/cm^2$ en cada año, se observó un claro incremento en la población. En el año 2018 este valor ocurrió el 4 de abril, en 2019 el 15 de abril, en 2020 en la primera semana de marzo, en 2021 el 6 de abril y en el año 2022 se tienen dos fechas una el 24 de febrero y otra en mayo. Este hallazgo nos permite tener un indicador sobre en qué momento es recomendable aplicar la técnica de control elegida.

- Como se mencionó previamente, la radiación solar total demostró ser una variable de gran impacto en el desarrollo del PHV. Se recomienda que en los lugares donde ya se tiene presencia, así como en lugares donde no, se instalen sensores de radiación, con ello poder darles seguimiento constante para que se logre una mayor certeza temporal sobre cuando se alcanzan los niveles de radiación relacionados con el incremento de la población de la plaga y se puedan tomar acciones oportunas.

- Con relación a las mediciones de temperatura disponibles, la temperatura de mayor impacto corresponde con la temperatura del suelo, lo cual concuerda con la literatura en donde se reporta que esta plaga se encuentra hasta 30 cm de profundidad. De acuerdo con la información analizada se observa el incremento de la plaga cuando se tienen temperaturas del suelo mayores a 23^oC (Fig. 52). Esto brinda una segunda oportunidad para tomar decisiones fundamentadas en una variable física que se puede medir.

- Aunque ya se mencionó, se recomienda ampliamente seguir realizando muestreos en todos los lotes, en zonas donde se tienen pequeñas incidencias e incluso donde aún no se tiene la presencia de plaga. Esto ayuda a evitar propagaciones así como para tomar medidas a tiempo, las cuales pueden ser tomadas a partir del ejemplo de Valle de Guadalupe.

Las pérdidas que se tienen por la presencia del PHV pueden ir desde una infestación muy leve con bajos niveles de pérdidas en materia y económicas, hasta infestaciones fuertes considerables, donde al no tomar las acciones pertinentes se pueden tener pérdidas del $100\,\%$ del cultivo, lo cual resulta en pérdidas económicas del orden de millones de dólares tal y como ocurrió en el estado de Sonora (Castillo et al., 2004). Para poder hablar de pérdidas económicas respecto a los niveles de infestación y a las acciones realizadas se debe considerar más factores. Un estudio más completo que indique el impacto económico

de tomar ciertas acciones, donde incluso el no hacer nada se considera como una acción, requiere de más información. Por ejemplo, Atallah et al. (2012) estudiaron el valor actual neto a lo largo de 25 años de vida de un viñedo, para examinar el impacto económico de la enfermedad del enrollamiento de la hoja (GLRD) en *Vitis vinifera* variedad Cabernet Frank, en los viñedos de Finger Lakes de Nueva York. Para lograr el objetivo emplearon información diversa como encuestas a los responsables de los viñedos, parámetros para escenarios de gestión de enfermedades, escenarios para evaluar el impacto del GLRD y análisis económicos para determinar los impactos económicos de la GLRD. Los resultados sugieren que, para minimizar las pérdidas potenciales debidas al GLRD, los gestores deberían prevenir la infección seleccionando vides certificadas y probadas contra el virus para la replantación, y que el control de la enfermedad debería basarse en los valores del nivel de infección, la reducción del rendimiento, la penalización del precio incurrido y la edad del viñedo.

Capítulo 5. Conclusiones y trabajo a futuro

En esta sección se presentan las conclusiones generales del presente trabajo de investigación, se describen las principales contribuciones, así como el trabajo futuro propuesto.

El sector agrícola está siendo afectado por diversos factores: crecimiento de la población, cambio climático, escasez de agua, plagas y enfermedades, por mencionar algunos. Se ha demostrado que es posible utilizar técnicas computacionales que ayudan a disminuir el efecto de alguno de los factores mencionados. En particular, el aprendizaje automático presenta oportunidades para descifrar, cuantificar y comprender diversos procesos en entornos agrícolas.

En la aplicación de técnicas de aprendizaje automático, en agricultura se identifican cuatro principales categorías: gestión del cultivo, gestión del agua, gestión del suelo y gestión del ganado. Dentro de la gestión del cultivo sobresalen las aplicaciones relacionadas con rendimiento de cultivo, detección de plagas y enfermedades, detección de malas hierbas, reconocimiento y calidad de los cultivos. Las plagas y enfermedades ponen en riesgo la seguridad alimentaria por medio de las grandes pérdidas que provocan, así como por el daño directo o indirecto que ocasionan a los humanos. En el presente trabajo se abordó el tema de predicción de plagas, empleando técnicas de aprendizaje de máquina basados en ejemplos.

El caso de estudio analizado corresponde con la predicción de la presencia y de los niveles de infestación del piojo harinoso de la vid en la región del Valle de Guadalupe. El estudiar esta plaga en territorio nacional es de vital importancia, ya que se han reportado pérdidas de hasta el $100\,\%$ de producción. Para lograr el objetivo se solicitó y fusionó información de vigilancia a nivel de campo, *i.e.*, muestreos constantes realizados por el grupo técnico del CESVBC, junto con información climatológica registrada diariamente por estaciones agroclimáticas.

Por medio de la metodología descrita se logró conocer los factores y parámetros que más impactan y propician el desarrollo de la plaga: el mes, la fenología, la radiación solar total y la temperatura del suelo fueron las características más importantes, las cuales que determinan en gran medida el incremento de la población de la plaga en el Valle de Guadalupe. Se encontró que en el mes de junio, principalmente si la vid se encuentra en crecimiento de baya, radiaciones superiores a los 600 cal/cm^2 y temperatura del suelo mayor a 23^oC son las condiciones donde se registraron los mayores picos de infestación en los predios analizados. El conocer esta información es relevante para que se pueden tomar diversas acciones, sean estas previas a las fechas mencionadas, así como la recomendación de la instalación de sensores de radiación en lugares estratégicos. De igual manera, si se desea realizar una consulta nueva,

por ejemplo, con las condiciones actuales se puede brindar una predicción obtenida por medio de los modelos entrenados y con base en ello decidir efectuar alguna acción.

Para poder predecir la presencia o el nivel de infestación se probaron dos modelos de clasificación basados en ejemplos: kNN y HSP. Ambos modelos obtuvieron métricas por encima del $73\,\%$ de correctas clasificaciones, sin embargo, considerando que el modelo HSP es libre de parámetros y que la predicción se realiza al momento de la consulta con base en la información almacenada, se considera el modelo más apto para el problema abordado. Este tipo de técnica se considera como una buena herramienta por su simplicidad y bajo costo computacional, así como por su explicabilidad; la asignación o la consulta se realiza por medio de comparación con los objetos cercanos obtenidos por el grafo HSP, estos objetos tienen la propiedad de ser similares a la consulta y diversos entre sí.

Como trabajo futuro se planea dar seguimiento al proyecto, es decir, se transmitirán los hallazgos obtenidos así como las recomendaciones a los interesados en un formato compacto y conciso. De igual manera, se ofrecerá la posibilidad de seguir alimentando las bases de datos existentes, con el objetivo de que cada vez los modelos sean más robustos, que se conozca más a detalle la dinámica del PHV y se puedan tomar las acciones pertinentes en regiones donde aún la incidencia es muy baja.

Adicionalmente, y gracias a los convenios generados durante la elaboración del proyecto, así como para sustentar la hipótesis planteada de que es posible aplicar la metodología descrita en otras plagas y otros cultivos, se planea implementar la metodología desarrollada para otro cultivo de gran importancia: el Nogal, en particular la plaga que se estudiará corresponde con el gusano barrenador; el cual presenta grandes pérdidas en diversos cultivos. Para lograr esto, se analizará información de muestreos y trampeos que se han llevado a cabo desde el año 2010 hasta la fecha en el estado de Sonora, siguiendo una metodología establecida por el comité de dicho estado.

Literatura citada

Agricultura-Senasica. 2021. Ficha técnica *Planococcus ficus (Signoret)* (hemiptera: Pseudococcidae) piojo harinoso de la vid. Dirección del Centro Nacional de Referencia Fitosanitaria.

Alsalam, B. H. Y., Morton, K., Campbell, D., y Gonzalez, F.. 2017. Autonomous uav with vision based on-board decision making for remote sensing and precision agriculture. En: 2017 IEEE Aerospace Conference..IEEE, pp. 1–12.

Altieri, M.. 2018. Agroecology: The Science of Sustainable Agriculture. CRC Press.

Aparecido, L. E., Rolim, G., Moraes, J., Costa, C., y Souza, P.. 2019. Machine learning algorithms for forecasting the incidence of coffea arabica pests and diseases. International Journal of Biometeorology, 64. doi: 10.1007/s00484-019-01856-1.

Atallah, S. S., Gómez, M. I., Fuchs, M. F., y Martinson, T. E.. 2012. Economic impact of grapevine leafroll disease on *Vitis vinifera* cv. *Cabernet franc* in finger lakes vineyards of new york. American Journal of Enology and Viticulture, 63(1), pp. 73–79. doi: 10.5344/ajev.2011.11055.

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., y Bochtis, D.. 2021. Machine learning in agriculture: A comprehensive updated review. Sensors, 21(11). doi: 10.3390/s21113758.

Bochtis, D. D., Sørensen, C. G., y Green, O.. 2012. A dss for planning of soil-sensitive field operations. Decision Support Systems, 53(1), pp. 66–75. doi: https://doi.org/10.1016/j.dss.2011.12.005.

Bourke, P. M. A.. 1970. Use of weather information in the prediction of plant disease epiphytotics. Annual Review of Phytopathology, 8(1), pp. 345–370. doi: 10.1146/annurev.py.08.090170.002021.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., y Sander, J.. 2000. Lof: Identifying density-based local outliers. ACM SIGMOD Record, 29(2), pp. 93–104. doi: 10.1145/335191.335388.

Burkov, A.. 2019. The Hundred-Page Machine Learning Book. Andriy Burkov.

Castillo, A. A. F., Blanco, J. L. M., Acosta, G. O., y Carrillo, J. L. M.. 2004. Control quimico de piojo harinoso *Planococcus ficus Signoret* (homoptera: Pseudococcidae) en vid de mesa. Agricultura Técnica en México, 30(1), pp. 101–105.

Chakraborty, S., Ghosh, R., Ghosh, M., Fernandes, C. D., Charchar, M. J., y Kelemu, S.. 2004. Weather-based prediction of anthracnose severity using artificial neural network models. Plant Pathology, 53(4), pp. 375–386. doi: https://doi.org/10.1111/j.1365-3059.2004.01044.x.

Chen, M., Brun, F., Raynal, M., y Makowski, D.. 2020. Forecasting severe grape downy mildew attacks using machine learning. PLOS ONE, 15(3), pp. 1–20. doi: 10.1371/journal.pone.0230254.

CICESE. 2022. Curso: El control biológico como base en el manejo integrado de plagas en la vid. Recuperado el 25 febrero de 2022 de https://youtu.be/bhxjt3qylba.

Coakley, S. M.. 1988. Variation in climate and prediction of disease in plants. Annual Review of Phytopathology, 26(1), pp. 163–181. doi: 10.1146/annurev.py.26.090188.001115.

Cocco, A., Pacheco da Silva, V. C., Benelli, G., Botton, M., Lucchi, A., y Lentini, A.. 2021. Sustainable management of the vine mealybug in organic vineyards. Journal of Pest Science, 94(2), pp. 153–185.

Conesa-Muñoz, J., Valente, J., Del Cerro, J., Barrientos, A., y Ribeiro, A.. 2016. A multi-robot sense-act approach to lead to a proper acting in environmental incidents. Sensors, 16(8). doi: 10.3390/s16081269.

Cormen, T. H.. 2009. Introduction to algorithms. MIT press.

Ding, C. y Peng, H.. 2003. Minimum redundancy feature selection from microarray gene expression data. En: Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003..pp. 523–528. doi: 10.1109/CSB.2003.1227396.

Durgabai, R., Bhargavi, P., et al.. 2018. Pest management using machine learning algorithms: a review. International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), 8(1), pp. 13–22.

ENA. 2019. Nota técnica encuesta nacional agropecuaria. Comunicado de prensa núm .481/20.

FAO. 2018. Los 10 elementos de la agroecología. Guía para la transición hacia sistemas alimentarios y agrícolas sostenibles. Roma, Italia, FAO, 24.

FAO. 2021. El estado mundial de la agricultura y la alimentación 2021. Lograr que los sistemas agroalimentarios sean más resilientes a las perturbaciones y tensiones. Roma, FAO. doi: https://doi.org/10.4060/cb4476es.

Ferentinos, K. P.. 2018. Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, pp. 311–318. doi: https://doi.org/10.1016/j.compag.2018.01.009.

Fikar, C.. 2018. A decision support system to investigate food losses in e-grocery deliveries. Computers Industrial Engineering, 117, pp. 282–290. doi: https://doi.org/10.1016/j.cie.2018.02.014.

Giusti, E. y Marsili-Libelli, S.. 2015. A fuzzy decision support system for irrigation and water conservation in agriculture. Environ. Model. Softw., 63(C), pp. 73–86. doi: 10.1016/j.envsoft.2014.09.020.

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., y Zhang, J.. 2020. Prediction of winter wheat yield based on multi-source data and machine learning in china. Remote Sensing, 12(2). doi: 10.3390/rs12020236.

Hasan, R. I., Yusuf, S. M., y Alzubaidi, L.. 2020. Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion. Plants, 9(10). doi: 10.3390/plants9101302.

Ishiguro, K. y Hashimoto, A.. 1991. Computer-based forecasting of rice blast epidemics in japan. En: International Rice Research Conference, Seoul (Korea Republic), 27-31 Aug 1990..IRRI.

Joyce, A., Hoddle, M., Bellows, T., y González, D.. 2001. Oviposition behavior of *Coccidoxenoides peregrinus*, a parasitoid of *Planococcus ficus*. Entomologia Experimentalis et Applicata, 98(1), pp. 49–57. doi: https://doi.org/10.1046/j.1570-7458.2001.00756.x.

Jurado, L. C.. 2020. Inifap promueve y evalúa el control biológico de plagas. Inovaciones para el campo, 1.

Kadiyala, M., Nedumaran, S., Singh, P., S., C., Irshad, M. A., y Bantilan, M.. 2015. An integrated crop model and gis decision support system for assisting agronomic decision making under climate change. Science of The Total Environment, 521-522, pp. 123–134. doi: https://doi.org/10.1016/j.scitotenv.2015.03.097.

Kaundal, R., Kapoor, A., y Raghava, G.. 2006. Machine learning techniques in disease forecasting: A case study on rice blast prediction. BMC bioinformatics, 7, pp. 485. doi: 10.1186/1471-2105-7-485.

Klem, K., Vánová, M., Hajslová, J., Ornerova, K., y Sehnalová, M.. 2007. A neural network model for prediction of deoxynivalenol content in wheat grain based on weather data and preceding crop. Plant, Soil and Environment, 53, pp. 421–429. doi: 10.17221/2200-PSE.

Li, D., Wang, R., Xie, C., Liu, L., Zhang, J., Li, R., Wang, F., Zhou, M., y Liu, W.. 2020. A recognition method for rice plant diseases and pests video detection based on deep convolutional neural network. Sensors, 20(3). doi: 10.3390/s20030578.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., y Bochtis, D. D.. 2018. Machine learning in agriculture: A review. Sensors, 18. doi: https://doi.org/10.3390/s18082674.

Liu, J. y Wang, X.. 2020. Tomato diseases and pests detection based on improved yolo v3 convolutional neural network. Frontiers in Plant Science, 11. doi: 10.3389/fpls.2020.00898.

Lu, J., Wu, D., Mao, M., Wang, W., y Zhang, G.. 2015. Recommender system application developments: A survey. Decision Support Systems, 74, pp. 12–32. doi: https://doi.org/10.1016/j.dss.2015.03.008.

Martínez, C.. 2018. La plaga que cambió el mapa del viñedo en España. Recuperado el 20 mayo de 2022 de https://www.mncn.csic.es/es/comunicacion/blog/la-plaga-que-cambio-el-mapa-del-vinedo-en-espana.

Mazzanti, S.. 2021. MRMR explained exactly how you wished someone explained to you. Recuperado el 10 febrero de 2022 de https://towardsdatascience.com/mrmr-explained-exactly-how-you-wished-someone-explained-to-you-9cf4ed27458b.

Meisner, M. H., Rosenheim, J. A., y Tagkopoulos, I.. 2016. A data-driven, machine learning framework for optimal pest management in cotton. Ecosphere, 7(3), pp. e01263. doi: https://doi.org/10.1002/ecs2.1263.

Meshram, V., Patil, K., Meshram, V., Hanchate, D., y Ramkteke, S.. 2021. Machine learning in agriculture domain: A state-of-art survey. Artificial Intelligence in the Life Sciences, 1, pp. 100010. doi: https://doi.org/10.1016/j.ailsci.2021.100010.

Nanni, L., Maguolo, G., y Pancino, F.. 2020. Insect pest image detection and recognition based on bio-inspired methods. Ecological Informatics, 57, pp. 101089. doi: https://doi.org/10.1016/j.ecoinf.2020.101089.

Navarro-Hellín, H., del Rincon, J. M., Domingo-Miguel, R., Soto-Valles, F., y Torres-Sánchez, R.. 2016. A decision support system for managing irrigation in agriculture. Computers and Electronics in Agriculture, 124, pp. 121–131. doi: https://doi.org/10.1016/j.compag.2016.04.003.

Oad, R., Garcia, L., Kinzli, K.-D., Patterson, D., y Shafike, N.. 2009. Decision support systems for efficient irrigation in the middle rio grande valley. Journal of Irrigation and Drainage Engineering, 135(2), pp. 177–185. doi: 10.1061/(ASCE)0733-9437(2009)135:2(177).

Perrier, A.. 2015. Feature importance in random forests. Recuperado el 10 agosto de 2022 de https://alexisperrier.com/datascience/2015/08/27/feature-importance-random-forests-gini-accuracy.html.

Pérez-Ariza, C., Nicholson, A., y Flores, M.. 2012. Prediction of coffee rust disease using bayesian networks. Proceedings of the 6th European Workshop on Probabilistic Graphical Models, 6, pp. 259–266.

Ramírez, I. C.. 2016. El desarrollo de la agricultura y el impacto que tendría en las finanzas públicas de México. Premio nacional de las finanzas públicas 2016.

Recio, B., Rubio, F., y Criado, J.. 2003. A decision support system for farm planning using agrisupport ii. Decision Support Systems, 36(2), pp. 189–203. doi: https://doi.org/10.1016/S0167-9236(02)00134-3.

Rodríguez-Moreno, V. M., Jiménez-Lagunes, A., Estrada-Avalos, J., Mauricio-Ruvalcaba, J. E., y Padilla-Ramírez, J. S.. 2020. Weather-data-based model: an approach for forecasting leaf and stripe rust on winter wheat. Meteorological Applications, 27(2), pp. e1896. doi: https://doi.org/10.1002/met.1896.

Russell, S. J. y Norvig, P.. 2009. Artificial Intelligence: a modern approach. 3a ed., Pearson.

SADER. 2018. Vino mexicano igual a excelencia. secretaría de agricultura y desarrollo rural. Recuperado el 5 septiembre 2022 de https://www.gob.mx/agricultura/articulos/vino-mexicano-igual-a-excelencia.

SAGARPA. 2011. Estudio estadístico sobre producción de uva en Baja California. Gobierno del estado Baja California.

Santos Valle, S. y Kienzle, J.. 2020. Agriculture 4.0 – agricultural robotics and automated equipment for sustainable crop production. Integrated Crop Management, 24 Roma, FAO.

Schütze, N. y Schmitz, G. H.. 2010. Occasion: New planning tool for optimal climate change adaption strategies in irrigation. Journal of Irrigation and Drainage Engineering, 136(12), pp. 836–846. doi: 10.1061/(ASCE)IR.1943-4774.0000266.

Shah, D., Trivedi, V., Sheth, V., Shah, A., y Chauhan, U.. 2021. Rests: Residual deep interpretable architecture for plant disease detection. Information Processing in Agriculture. doi: https://doi.org/10.1016/j.inpa.2021.06.001.

Sharma, L., Gonçalves, F., Oliveira, I., Torres, L., y Marques, G.. 2018. Insect-associated fungi from naturally mycosed vine mealybug *Planococcus ficus (Signoret)* (hemiptera: Pseudococcidae). Biocontrol Science and Technology, 28(2), pp. 122–141. doi: 10.1080/09583157.2018.1428733.

SIAP. 2021. Anuario estadístico de la producción agrícola. Recuperado el 10 agosto 2022 de https://nube.siap.gob.mx/cierreagricola/.

Soysal, M., Bloemhof-Ruwaard, J., y van der Vorst, J.. 2014. Modelling food logistics networks with emission considerations: The case of an international beef supply chain. International Journal of Production Economics, 152, pp. 57–70. Sustainable Food Supply Chain Management. doi: https://doi.org/10.1016/j.ijpe.2013.12.012.

Talamantes, A. y Chavez, E.. 2022. Instance-based learning using the half-space proximal graph. Pattern Recognition Letters, 156, pp. 88–95.

Talamantes A., A.. 2021. Aprendizaje basado en ejemplos mediante el grafo de semi-espacios proximales. Tesis de Maestría. Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California. 68 hojas.

Ting, S., Tse, Y., Ho, G., Chung, S., y Pang, G.. 2014. Mining logistics data to assure the quality in a sustainable food supply chain: A case in the red wine industry. International Journal of Production Economics, 152, pp. 200–209. Sustainable Food Supply Chain Management. doi: https://doi.org/10.1016/j.ijpe.2013.12.010.

Trebuna, P., Halcinová, J., Fil'o, M., y Markovic, J.. 2014. The importance of normalization and standardization in the process of clustering. 01..pp. 381–385. doi: 10.1109/SAMI.2014.6822444.

Tripathy, A. K., Adinarayana, J., Sudharsan, D., Merchant, S. N., Desai, U. B., Vijayalakshmi, K., Raji Reddy, D., Sreenivas, G., Ninomiya, S., Hirafuji, M., Kiura, T., y Tanaka, K.. 2011. Data mining and wireless sensor network for agriculture pest/disease predictions. En: 2011 World Congress on Information and Communication Technologies..pp. 1229–1234. doi: 10.1109/WICT.2011.6141424.

Valladolid, M., Ruiz, J., Méndez, S., y Velasco, L.. 2018. Manual básico de viticultura en Baja California sistema de producción anual Ensenada, Baja California. Universidad Autónoma de Baja California.

Wang, H. y Ma, Z.. 2011. Prediction of wheat stripe rust based on support vector machine. En: 2011 Seventh International Conference on Natural Computation..IEEE, Vol. 1, pp. 378–382.

Wenkel, K.-O., Berg, M., Mirschel, W., Wieland, R., Nendel, C., y Köstner, B.. 2013. Landcare dss – an interactive decision support system for climate change impact assessment and the analysis of potential agricultural land use adaptation strategies. Journal of Environmental Management, 127, pp. S168–S183. doi: https://doi.org/10.1016/j.jenvman.2013.02.051.

Xenakis, A., Papastergiou, G., Gerogiannis, V. C., y Stamoulis, G.. 2020. Applying a convolutional neural network in an iot robotic system for plant disease diagnosis. En: 2020 11th International Conference on Information, Intelligence, Systems and Applications..pp. 1–8. doi: 10.1109/IISA50023.2020.9284356.

Xiao, Q., Li, W., Kai, Y., Chen, P., Zhang, J., y Wang, B.. 2019. Occurrence prediction of pests and diseases in cotton on the basis of weather factors by long short term memory network. BMC Bioinformatics, 20. doi: 10.1186/s12859-019-3262-y.

Zepeda-Jazo, I.. 2018. Manejo sustentable de plagas agrícolas en México. Agricultura, sociedad y desarrollo, 15, pp. 99 – 108.

Zhai, Z., Martínez, J. F., Beltran, V., y Martínez, N. L.. 2020. Decision support systems for agriculture 4.0: Survey and challenges. Computers and Electronics in Agriculture, 170, pp. 105256. doi: https://doi.org/10.1016/j.compag.2020.105256.

Zhang, S., Li, X., Zong, M., Zhu, X., y Wang, R.. 2017. Efficient knn classification with different numbers of nearest neighbors. IEEE Transactions on Neural Networks and Learning Systems, PP, pp. 1–12. doi: 10.1109/TNNLS.2017.2673241.

Anexo A

Mapas de calor donde se aprecia la distribución temporal de la incidencia del PHV en las comunidades de los cuatro modelos predictivos propuestos.

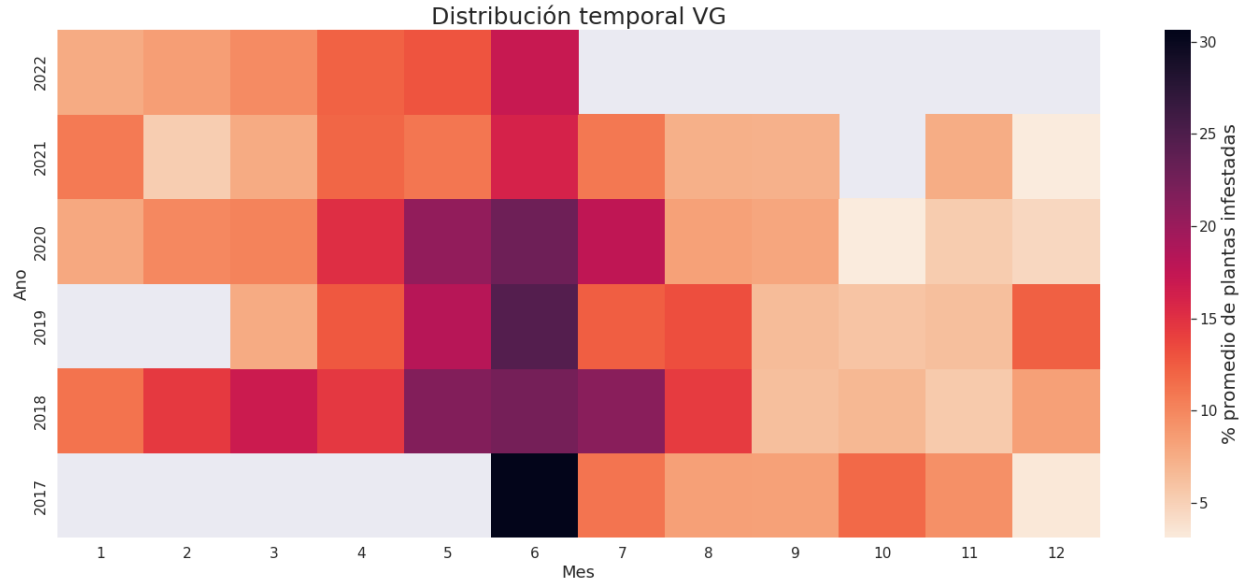

Figura 47. Distribución temporal en Valle de Guadalupe.

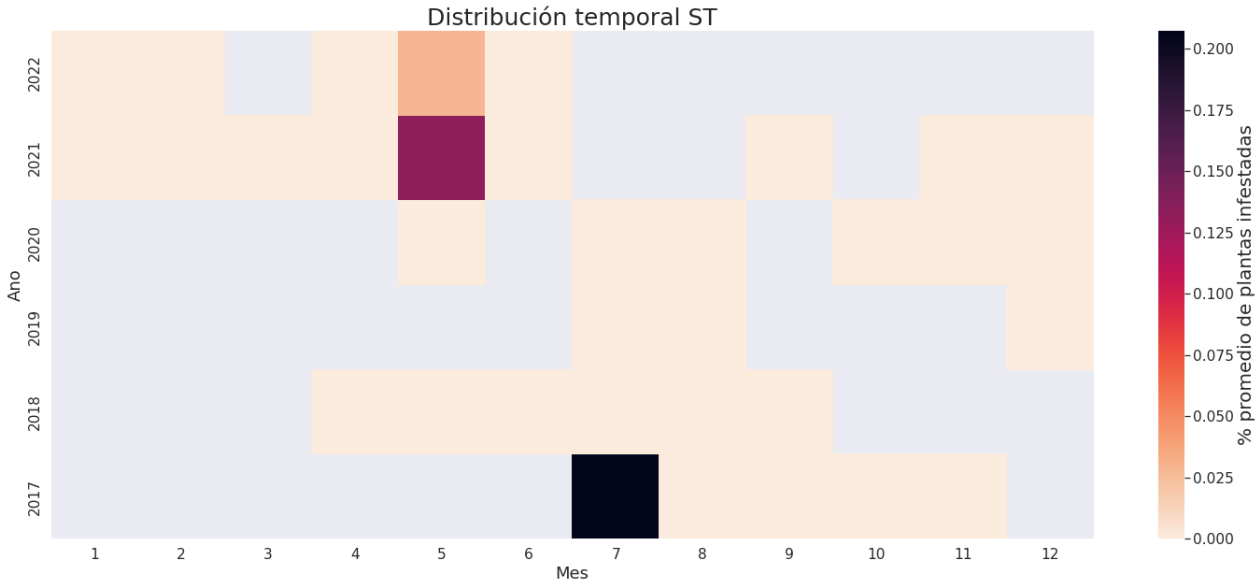

Figura 48. Distribución temporal en Santo Tomás.

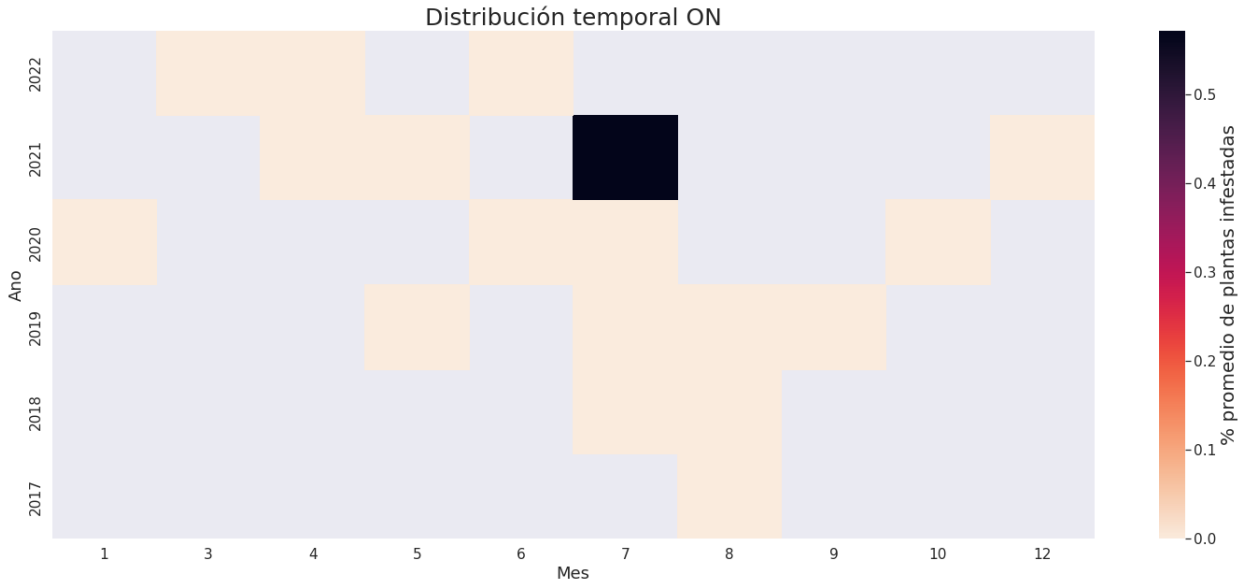

Figura 49. Distribución temporal en Ojos Negros.

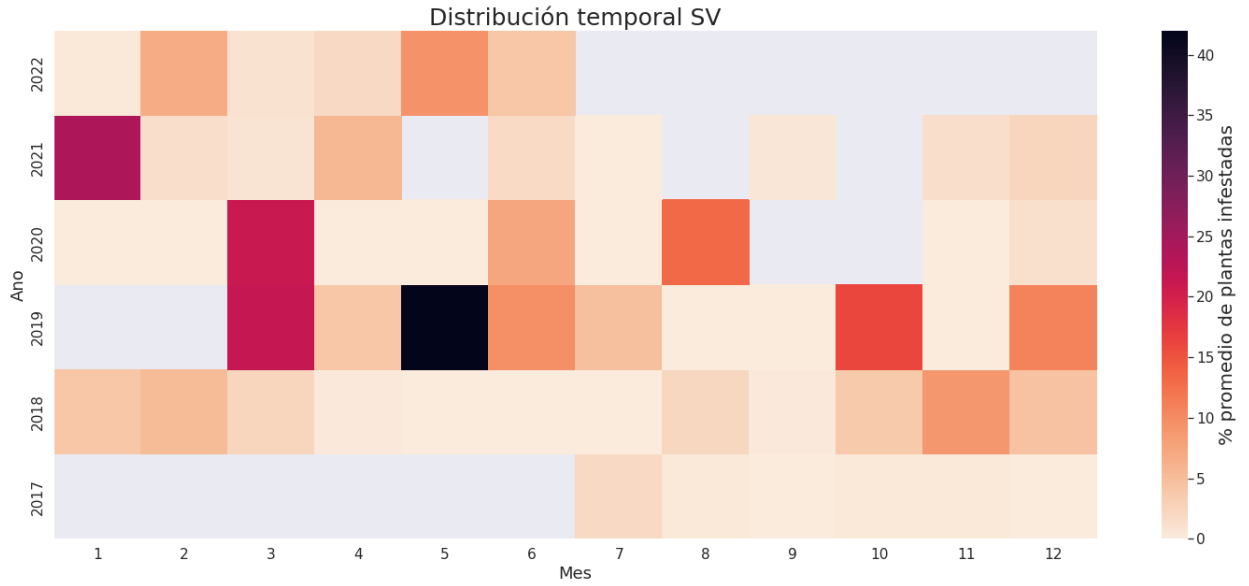

Figura 50. Distribución temporal en San Vicente.

Anexo B

Imágenes de diversas variables seleccionadas respecto a la cantidad de piojos encontrados en los muestreos para el modelo 1 de Valle de Guadalupe.

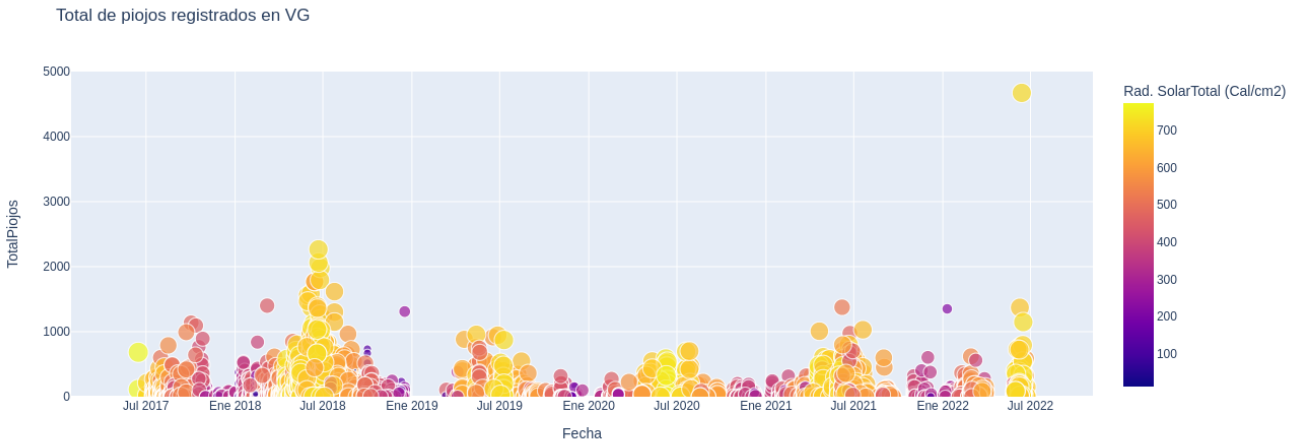

Figura 51. Número de piojos encontrados en el periodo de muestro en comparación con la radiación solar total.

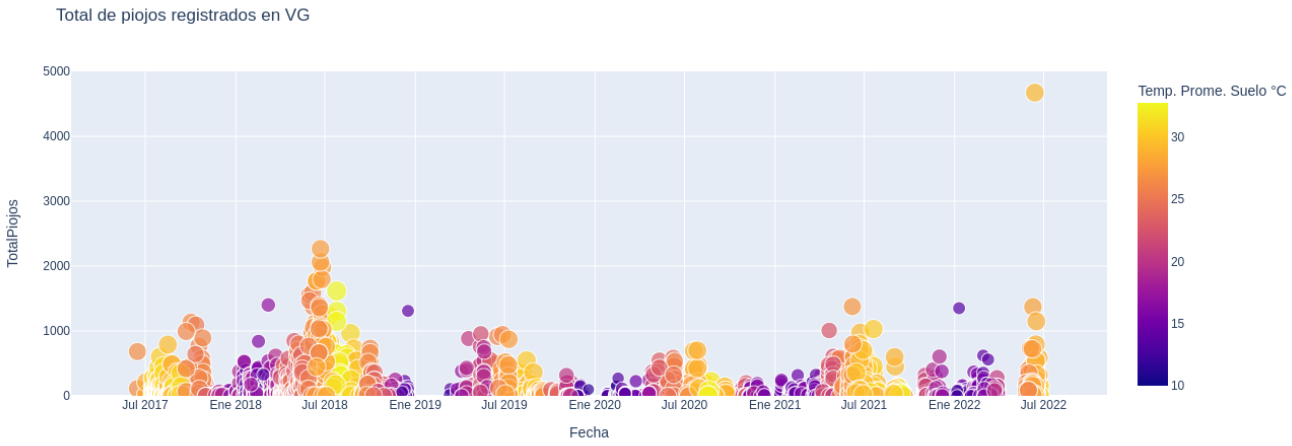

Figura 52. Número de piojos encontrados en el periodo de muestro en comparación con la temperatura de suelo.

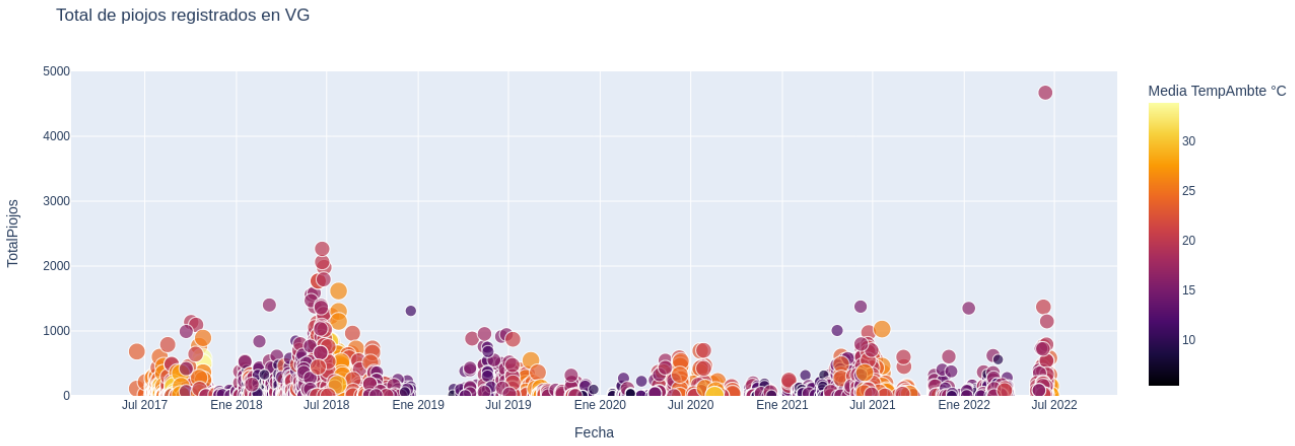

Figura 53. Número de piojos encontrados en el periodo de muestro en comparación con la temperatura ambiental promedio.

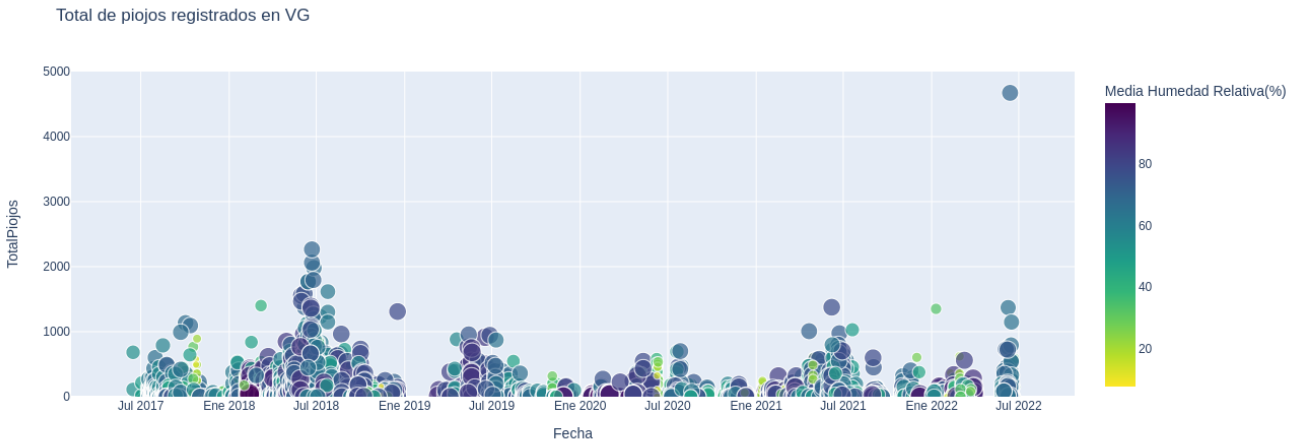

Figura 54. Número de piojos encontrados en el periodo de muestro en comparación con la humedad relativa promedio.

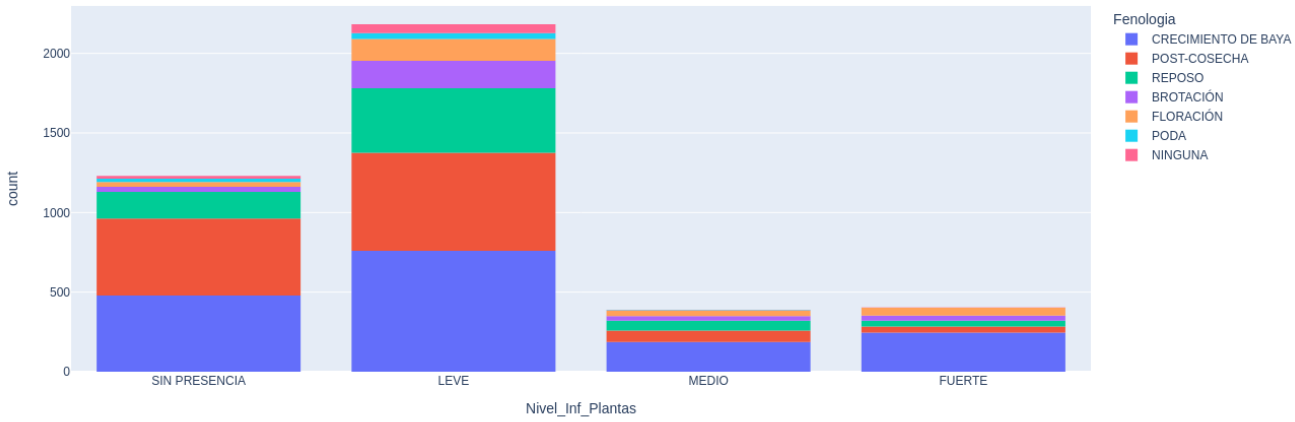

Figura 55. Niveles de infestación en Valle de Guadalupe señalando la fenología.