# UNIVERSIDAD PERUANA UNIÓN

ESCUELA DE POSGRADO

Unidad de Posgrado de Posgrado de Ingeniería y Arquitectura


## Modelo de predicción de la plaga Burkholderia Glumae en cultivos de arroz usando Machine Learning e Interpolación Espacial

Tesis para obtener el Grado Académico de Maestro en Ingeniería de Sistemas con mención en Ingeniería de Software.

**Autor:**

Joel Pérez Suárez

**Asesor:**

Mg. Nemias Saboya Ríos

Lima, febrero de 2022

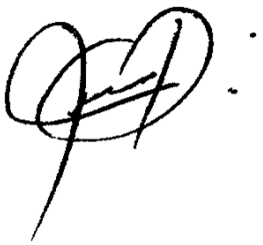
# DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo Nemias Saboya Ríos, docente de la Unidad de Posgrado de Ingeniería y Arquitectura, Escuela de Posgrado de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: **"MODELO DE PREDICCIÓN DE LA PLAGA BURKHOLDERIA GLUMAE EN CULTIVOS DE ARROZ USANDO MACHINE LEARNING E INTERPOLACIÓN ESPACIAL"** del autor Joel Pérez Suárez tiene un índice de similitud de 15% verificable en el informe del programa Turnitin, y fue realizada en la Universidad Peruana Unión bajo mi dirección.

En tal sentido asumo la responsabilidad que corresponde ante cualquier falsedad u omisión de los documentos como de la información aportada, firmo la presente declaración en la ciudad de Lima, a los 3 días del mes de febrero del año 2022.



Nemias Saboya Ríos

# ACTA DE SUSTENTACIÓN

**ACTA DE SUSTENTACIÓN DE TESIS DE MAESTRO(A)** 315

En Lima, Ñaña, Villa Unión, a 03 días del mes de febrero del año 2022, siendo las 10:30 a. m, se reunieron en la modalidad online sincrónica, bajo la dirección del Señor Presidente del Jurado: Mg. Immer Elías Cuellar Rodríguez, el secretario: Mg. Danny Lévano Rodríguez, los demás miembros: Mg. Cynthia Carol Acuña Salinas y el Mg. Abel Angel Sullon Macalupu y el asesor: Mg. Nemias Saboya Ríos, con el propósito de administrar el acto académico de sustentación de Tesis de Maestro(a) titulada: "Modelo de predicción de la plaga Burkholderia Glumae en cultivos de arroz usando Machine Learning e Interpolación Espacial" del Bachiller/Licenciado(a) Joel Pérez Suárez

Conducente a la obtención del Grado Académico de Maestro(a) en: Ingeniería de Sistemas

(Nomenclatura del Grado Académico)

con Mención en Ingeniería de Software El Presidente inició el acto académico de sustentación invitando al candidato hacer uso del tiempo determinado para su exposición. Concluida la exposición, el Presidente invitó a los demás miembros del Jurado a efectuar las preguntas, cuestionamientos y aclaraciones pertinentes, los cuales fueron absueltos por el candidato. Luego se produjo un receso para las deliberaciones y la emición del dictamen del Jurado.

Posteriormente, el Jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Bachiller/Licenciado (a): Joel Pérez Suárez

| CALIFICACIÓN | ESCALAS | | | Mérito |
|---|---|---|---|---|
| | Vigesimal | Literal | Cualitativa | |
| Aprobado | 17 | B+ | Con nominación de Muy Bueno | Sobresaliente |

*(*) Ver parte posterior*

Finalmente, el Presidente del Jurado invitó al candidato a ponerse de pie, para recibir la evaluación final. Además, el Presidente del Jurado concluyó el acto académico de sustentación, procediéndose a registrar las firmas respectivas.

Presidente

Secretario

Asesor

Miembro

Miembro

Bachiller/Licenciado(a)

**Agradecimiento**

A Dios, por ser la fuente de la sabiduría.

A mi esposa Deisy Diana Díaz Salcedo por seguir creyendo en mí.

A mi asesor Mg. Nemias Saboya Ríos por su entrega, perseverancia y compromiso con este trabajo de investigación.

Al Mg. Abel Ángel Sullón Macalupú por su apoyo en el desarrollo de la investigación.

**Dedicatoria**

A mi esposa Deisy Diana Díaz Salcedo y mis hijos (Janice Eliane y Joe David).

A mis padres Wilmer A. Pérez Rodas y Azucena Suárez Flores.

A mi hermano Edwin Salvador Pérez Suárez

A mis segundos padres Flavio Díaz Banda y Lidia Salcedo Muñoz.

# ÍNDICE

# Modelo de predicción de la plaga Burkholderia Glumae en cultivos de arroz usando Machine Learning e Interpolación Espacial

Joel Perez-Suarez [1], *Nemias Saboya [2] and A. Angel Sullon [3]

[1] Unidad de Posgrado de Ingeniería y Arquitectura, Universidad Peruana Unión, Carretera Central Km 19.5 Lurigancho, Lima, Perú
joel.perez@upeu.edu.pe
[2]Escuela Profesional de Ingeniería de Sistemas, Universidad Peruana Unión, Carretera Central Km 19.5 Lurigancho, Lima, Perú
saboya@upeu.edu.pe
[3]Escuela Profesional de Ingeniería de Sistemas, Universidad Peruana Unión, Filial Juliaca, Carretera salida a Arequipa Km 6 Chullunquiani, Juliaca, Perú
angeli@upeu.edu.pe

**Abstract**. En la actualidad, la agricultura, en especial el cultivo de arroz está siendo afectado por los constantes cambios climáticos, esto origina que algunos patógenos sean favorecidos alterando su producción, y a su vez generando pérdidas económicas. El estudio tuvo el propósito de elaborar un modelo de Machine Learning para predecir la aparición de la plaga Burkholderia Glumae en cultivos de arroz en la región de San Martín, Perú. En la exploración de los datos se usó la técnica de interpolación espacial IDW, para obtener datos de temperatura y precipitación. El estudio aplicó una serie de algoritmos supervisados. Entre estos, el Random Forest Classifier (RFC) fue el que obtuvo el máximo valor con un accuracy de 88%. También se creó un aplicativo donde se visualiza la predicción de la plaga Burkholderia Glumae en una zona determinada de la región.

**Keywords:** Machine Learning, plaga de arroz, Burkholderia Glumae, Random Forest Classifier, Interpolación Espacial.

## 1 Introduction

El arroz es uno de los componentes principales en la alimentación, de la población peruana, por esta razón, los agricultores de diferentes partes del país consideran de suma importancia la siembra y cosecha de este producto, sin embargo, existen dificultades para su óptima producción. Por ejemplo, la presencia de plagas, enfermedades y malezas, estas situaciones perjudican durante el proceso de crecimiento y desarrollo, generando que los costos de producción se eleven [1],[2].

En la actualidad, la agricultura de China, Corea, Colombia, Estados Unidos, Israel, Brasil y entre otros países, previenen riesgos durante el proceso de siembra y cosecha del arroz, usando tecnologías que pronostican eventos o situaciones que conlleven a pérdidas económicas. Estas tecnologías están basadas en la explotación de decenas de datos históricos apoyado con el desarrollo de hardware y software inteligente [3],[4],[5], [6]. Por otro lado, existen diversos estudios orientados a la agricultura, como el estudio realizado por Dhaya, donde estudió las enfermedades de las plantas en diversas partes que las componen utilizando imágenes y algoritmos de machine learning obteniendo un accuracy del 96% [7], de la misma manera el estudio realizado por Shakya utiliza algoritmos inteligentes de clasificación y escala espacial con el propósito de realizar un análisis de productividad del suelos para los cultivos de maíz y garbanzo, concluyendo que el estudio fue eficiente ya que los cultivos fueron más productivos [8].

Uno de los problemas recurrentes de la agricultura, dependiendo del tipo de producción es ocasionada por los cambios climáticos extremos (Temperatura, precipitación, humedad y entre otros.) porque a través de estos cambios se incrementan las probabilidades de la presencia de las plagas y las enfermedades[6]; asimismo, con el cambio climático en el contexto actual se pronostica que se será mayor a mediados del siglo XXI [9],[10].

En el Perú, el cultivo del arroz no es ajeno a los problemas fitosanitarios que afectan a la plantación y al producto en sí. Las plagas que se encontraron con más frecuencia son: Rhizoctonia solani, Gaeumannomvces sp y Nakataea sp, estos afectan al tallo y a la hoja de la planta de arroz descomponiéndola. Las Pyricularia oryzae, producen manchas foliares. Las Steneotarfsonemus spinki son ácaros blancos de la panoja de arroz y la bacteria Burkholderia glumae se encuentra en la panícula de arroz [9]. Esta última fue la que más se encontró en el cultivo de arroz, que durante el 2014, afectó la cosecha de arroz en la provincia de Tumbes, ocasionando la disminución de la producción en 40% y 60% [9], en Piura, Lambayeque y San Martín también fueron afectados, pues su producción se ha retraído entre 15% y 75% [11].

La investigación tuvo el propósito de elaborar un modelo de Machine Learning que contribuya en la predicción de la aparición de la plaga Burkholderia Glumae en el cultivo de arroz en la región de San Martin, Perú, considerando los factores climáticos.

La región San Martín, una de las más importantes en la producción de arroz del Perú con 14 mil agricultores[2], donde se realizan actividades agropecuarias que son consideradas en el marco del crecimiento sectorial, en tal sentido, es de gran preocupación para los agricultores considerar este aspecto durante la planificación y proyección del cultivo de arroz, pues en muchos casos, se vieron obligados a interrumpir por distintos factores como las inoportunas situaciones climáticas o de plagas que afectan la situación económica generando la disminución de la producción agropecuaria [2].

### 1.1 Cultivo de arroz

"El arroz es el alimento básico de 17 países de Asia y del Pacífico, de ocho países de África, de siete países de América Latina y del Caribe y de uno del Cercano Oriente" [12]. En el Perú este cereal junto con la papa ocupa las mayores áreas de cultivo, agrupan a la mayor cantidad de productores y aportan en mayor magnitud al Valor Bruto de la Producción (VBP) agrícola[12]. Asimismo, el arroz y la papa juntos representan aproximadamente el 26% de PBI agropecuario[1]. El departamento de San Martín ocupa el primer lugar en sembrío de este cereal, ya que en la campaña 2020-2021 en promedio de 101,890 hectáreas han sido sembradas[13].

### 1.2 Plagas de arroz

**Burkholderia Glumae**. También se le denomina como el Añublo Bacteriano de la panícula del arroz. Esta plaga causa la descomposición de granos y plántulas. Asimismo, se transmite principalmente por la semilla, además, muestra síntomas en la etapa de floración y luego se despliega por las espigas generando panículas de color café, debido a un proceso de clorosis. En tal sentido, si aún las plantas se mantienen erguidas es debido a la falta de peso en las ramas. También se ha comprobado que la temperatura alta favorece al patógeno, especialmente por la noche con una alta humedad relativa y con precipitaciones frecuentes [14].

**Bipolaris Oryzae.** Esta plaga también se le conoce como Mancha Parda del arroz o Mancha Marrón, y es común en zonas tropicales con suelos infértiles. El hongo es favorecido con temperaturas entre 25° y 34° [15].

**Bulkholderia Gladioli.** Denominado como un patógeno de plantas, además se caracteriza por poseer formas de nichos ecológicos en el suelo, plantas y las vías respiratorias humanas, también se dice que tiene mucha influencia en enfermedades pulmonares[16].

### 1.3 Factores climáticos

Las altas y bajas temperaturas influyen en el desarrollo del macollamiento, la formación de espiguilla y maduración, retrasando el crecimiento de la planta de arroz [17]. Asimismo, otro factor es la precipitación, esto se caracteriza por la cantidad de agua que cae de la superficie atmosférica, puede ser liquido o sólido. En estado líquido comprende la lluvia o llovizna y en estado sólido nieve, granizo o escarcha [18]. Además, la unidad de medida es en milímetros(mm), donde 1 mm es un litro en un metro cuadrado [18]. La lluvia consiste en gotas de agua muy fuertes, cuyo diámetro esta desde los 5mm [18]. La llovizna o garua son pequeñas gotas de agua cuyo diámetro es de 0.1 a 0.5 mm [18].

### 1.4 Interpolación espacial de datos

La interpolación espacial es un procedimiento matemático o geo estadístico para calcular un punto desconocido de una superficie, respecto a otros puntos vecinos cercanos de la misma región, entre las técnicas más conocidas se encuentran Kriging Interpolation, Polynomial Interpolation, Inverse Distance Weighted Interpolation (IDW) y Triangulated Irregular Network(TIN) [19]. Además, es considerada una técnica determinística porque se basa en la matemática, su fórmula es:

$$z_p = \frac{\sum_i^n (Z_i / d(p1,p2)^P)_i}{\sum_i^n (1/d(p1,p2)^P)_i} \quad (1)$$

Donde (1):

z, es el parámetro conocido de los vecinos cercanos (ejemplo: temperatura).

d(p1, p2), es la distancia entre dos puntos.

p, es la potencia, por lo general es 1 ó 2 [20].

### 1.5 Algoritmos de Machine Learning(ML)

Los algoritmos de ML se agrupan en 4 componentes principales. En primer lugar, aprendizaje supervisado, donde por lo general se usan algoritmos de clasificación y regresión. En segundo lugar, aprendizaje no supervisado, donde se usan algoritmos de agrupación, asociación y entre otros. En tercer lugar, aprendizaje semi supervisado, combina algoritmos supervisados y no supervisados. En cuarto lugar, aprendizaje por refuerzo [21], [22].

En el estudio se utilizó algoritmos de clasificación por las características de las variables. Estas fueron: Decisión Tree(DT) [23], Extreme Gradient Boosting(XGB) [24], Random Forest (RF) [25], Regresión Logistica (LRN), Linear Discriminant Analysis LDA, Support Vector Machines (SVM), Stochastic Gradient Descent (SGD), KNearest Neighbors (KNN), Gaussian Naive Bayes (GNB) y Neural network models (NNM).

## 2 Metodología

En el procedimiento metodológico se utilizó las buenas prácticas de CRISP-DM y otros estudios [26], [22], siendo este una de las metodologías más importantes que orientaron el desarrollo de un modelo para Machine Learning. Esto comprendió de 4 fases: comprensión de los datos, preparación de los datos,

modelado y despliegue del modelo. Cada fase, así como las actividades se detallan en la siguiente imagen. Ver fig. 1.

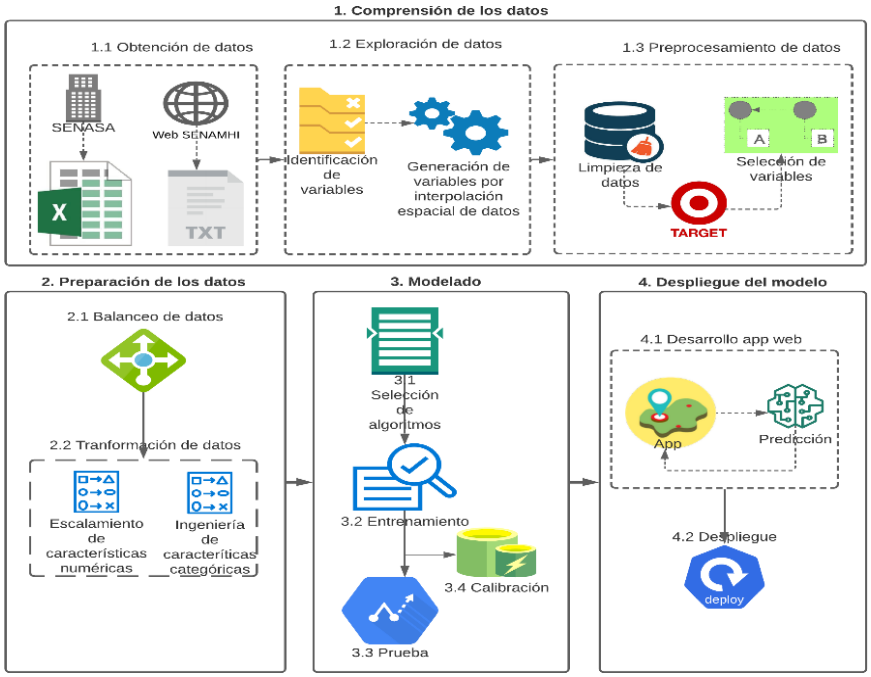



**Fig. 1.** Modelo predictivo de ML de predicción de la plaga Burkholderia Glumae

## 2.1 Comprensión de los datos

### Obtención de los datos.

Los datos se obtuvieron de 2 fuentes: La primera, fue solicitada al Servicio Nacional de Sanidad Agraria de San Martín (SENASA) [27]; estos proporcionaron datos de las plagas identificadas en los cultivo de arroz entre los años 2009 y 2018. La segunda, se obtuvo del Servicio Nacional de Meteorología e Hidrología del Perú(SENAMHI) [28], según las plantas meteorológicas de la región San Martín de los años de 1964 hasta el 2019 a través de algoritmos de Web Scraping con Python y se descargó los datos de forma gratuita de la página oficial [29].

### Exploración de los datos.

Se realizó 2 tareas principales: La identificación de variables y la generación de variables por interpolación espacial de datos. En la primera, se seleccionó las variables de la data obtenida del SENASA y SENAMHI, esas variables fueron seleccionadas bajo el criterio de los resultados del análisis exploratorios y descriptivos realizados con R_studio. En la segunda, se generaron variables adicionales que se consiguieron bajo el análisis de la interpolación espacial de datos utilizando la técnica IDW; obteniendo finalmente 31 variables que fueron consideradas para el desarrollo del modelo, estas variables y su descripción se detallan en la tabla.

**Table 1.** Variables consideradas para el desarrollo del modelo

| N° | Variables | Descripción | N° | Variables | Descripción |
|---|---|---|---|---|---|
| 1 | Vegetal_n_comun | Nombre común de la planta a cosechar | 17 | temp_max l_3 | Temperatura máxima del tercer día |

| | | | | | |
|---|---|---|---|---|---|
| 2 | Fecha_recepcion | Fecha de recepción de la muestra | 18 | temp_min l_3 | Temperatura mínima del tercer día |
| 3 | Tipo_muestra | Describe de que parte de la planta fue sacado la muestra(espiga, tallo, grano, hoja, otros) | 19 | temp_prom l_3 | Temperatura promedio del tercer día |
| 4 | Departamento | Departamento donde se realizó la muestra | 20 | precip l_2 | Precipitación del segundo día |
| 5 | Provincia | Provincia donde se realizó la muestra | 21 | temp_max l_2 | Temperatura máxima del segundo día |
| 6 | Distrito | Distrito donde se realizó la muestra | 22 | temp_min l_2 | Temperatura mínima del segundo día |
| 7 | Motivo | Describe la ausencia o presencia de la plaga en el cultivo muestreado | 23 | temp_prom l_2 | Temperatura promedio del segundo día |
| 8 | n_cientifico_plaga | Nombre científico de la plaga | 24 | precip l_1 | Precipitación del primer día |
| 9 | n_comun_plaga | Nombre común de la plaga | 25 | temp_max l_1 | Temperatura máxima del primer día |
| 10 | Latitud | Describe la latitud del sembrío | 26 | temp_min l_1 | Temperatura mínima del primer día |
| 11 | Longitud | Describe la longitud del sembrío | 27 | temp_prom l_1 | Temperatura promedio del primer día |
| 12 | Altitud | Altitud del sembrío | 28 | precip l_0 | Precipitación del día de la recepción de la muestra |
| 13 | Área_semb_sem | Área de sembrío en hectáreas | 29 | temp_max l_0 | Temperatura máxima del día de la recepción de la muestra |
| 14 | Área_afec_cul | Área afectada en todo el sembrío | 30 | temp_min l_0 | Temperatura mínima del día de la recepción de la muestra |
| 15 | Compute_0054 | Describe el porcentaje o hectáreas afectadas en el sembrío | 31 | temp_prom l_0 | Temperatura promedio del día de la recepción de la muestra |
| 16 | Precip l_3 | Precipitación del tercer día | | | |

En la interpolación de los datos se consideró 34 estaciones meteorológicas de la región San Martín, estas estaciones se encontraban en zonas cercanas de cultivos de arroz (ver fig. 2); donde la representación del triángulo de color rojo son las estaciones y los cuadrados de color verde son los sembríos de arroz, esta información fue de utilidad ya que permitió validar la aplicabilidad de la interpolación espacial.

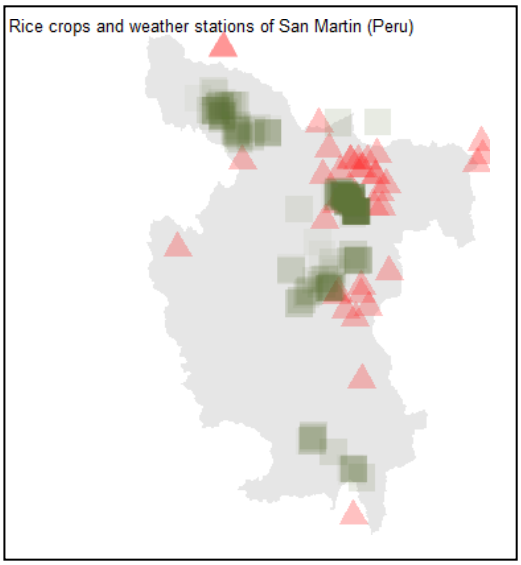



**Fig. 2.** Estaciones meteorológicas y sembríos de arroz

Por otro lado, la información que se obtuvo del SENASA fue de los últimos 10 años y estuvo distribuida de la siguiente manera: San Martin con 155 registros que representa el 40.365%, Rioja con 57 que representa el 14.844%, Moyobamba con 56 que presenta un 14.583%, Bellavista con 37 que representa un 9.635 %, Tocache con 30 que representa un 7.812%, Picota con 22 que es 5.729%, Mariscal Cáceres con 11 que es 2.865%, Alto Amazonas con 9 que es 2.344%, Huallaga con 5 que es 1.302% y el Dorado con 2 que hacen el 0.521%, donde toda esta información representa un total de 384 registros respecto a las plagas del cultivos de arroz, dicha cantidad se utilizó para la interpolación considerando un Inverso a la Distancia Ponderada (IDP) de 2 [19] porque se requería información de temperatura y precipitación, algunos ejemplos se representan en la figura 3. Por otro lado, se identificó que el 51.359% presentaron la plaga Burkholderia glumae, el 18.478% Bipolaris oryzae, el 17.935 % Burkholderia gladioli y 12.228% otras plagas. En la investigación, se consideró a la plaga de Burkholderia glumae representadas en las 6 provincias con más registros recolectados.

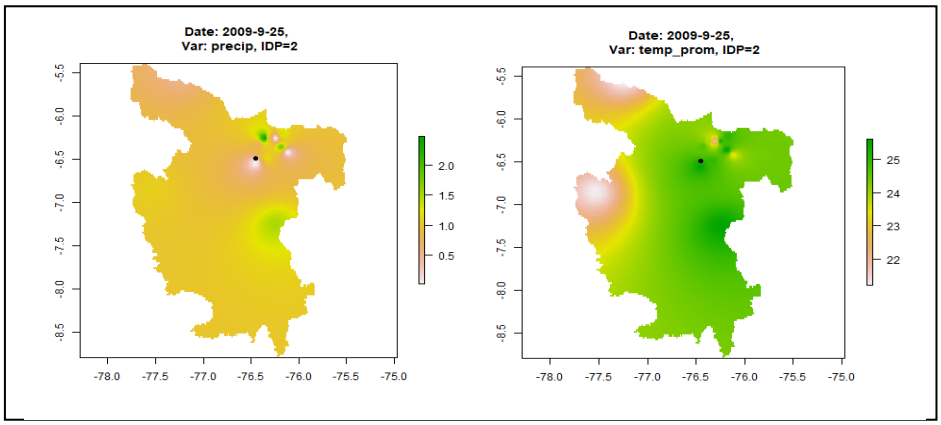



**Fig. 3.** Resultado de la interpolación que representa la precipitación y temperatura.

Asimismo, en la fig. 3 que representa a la región San Martín, el punto negro en el mapa simboliza el valor de la temperatura y precipitación, donde el color verde oscuro denota una temperatura mayor, el color blanco una temperatura menor, en la precipitación el color verde denota lluvias intensas y de color blanco ausencia de lluvias.

**Pre procesamiento de datos.**
En esta etapa, se realizó 3 tareas principales: limpieza de datos, definición de objetivos(target) y selección de variables. En la limpieza de datos, se encontró 384 registros iniciales posteriormente al proceso de interpolación quedaron 381, y se eliminaron 49 nulos, donde se tuvo un total de 332 registros. En la definición de objetivos se identificó 2 variables claves, "área de afectación del cultivo y área sembrada", para el área de afectación de cultivo fue necesario una conversión de los valores a porcentaje con el objetivo de normalizar la información e identificar en porcentaje el área que afecta la plaga Burkholderia Glumae en un sembrío y a su vez para poder identificar el estadístico más adecuado(mediana) para representar la presencia o ausencia de una plaga.

Por último, se agregó la variable objetivo llamado "TARGET"; donde se clasificó a los valores menores e iguales al valor de la mediana con ausencia de plagas (p-) y mayores a la mediana con presencia de plagas (p+); ver script.

```
dfo1['res dfo1['TARGET']=np.where( dfo1['response_var'] <= 20 , 'p-' , 'P+')
```

En la selección de variables, se utilizó la prueba estadística de Coeficiente de Correlación de Pearson con el objetivo de seleccionar a las variables más representativas que aportarían al modelo [30]. Estas variables pasaron por un proceso exploratorio y de normalización, porque todas las variables son de tipo cuantitativas. En la tabla 2 se mencionan las variables que se seleccionaron considerando su valor correlacional más alto.

**Table 2.** Variables seleccionadas para el modelo de ML

| N° | Variables | Descripción | Corr. |
|---|---|---|---|
| 1 | Latitud | Latitud de sembrío | -0.122602 |
| 2 | Longitud | Longitud de sembrío | 0.125838 |
| 3 | Precip l_0 | Precipitación del día de la recepción de la muestra | 0.212568 |
| 4 | Temp_max l_0 | Temperatura máxima del día de la recepción de la muestra | 0.092151 |
| 5 | Temp_min l_0 | Temperatura mínima del día de la recepción de la muestra | 0.081526 |
| 6 | Temp_max l_2 | Temperatura máxima del segundo día | 0.058687 |
| 7 | Temp_min l_2 | Temperatura mínima del segundo día | -0.047667 |
| 8 | Precip l_2 | Precipitación del segundo día | -0.116582 |
| 9 | TARGET | Variable objetivo | 1.000000 |

## 2.2 Preparación de los datos

**Balanceo de datos**
Se utilizó la técnica de RandomOverSampler donde se considera la sobre muestra de la clase minoritaria y se elige muestras aleatorias con reemplazo, esto con el objetivo de igualar los registros del cultivo de arroz de plagas positivas y negativas, como lo detalla la tabla 3, donde denota desigualdad de registros en las muestras.

**Table 3.** Análisis de balanceo de datos

| TARGET | Cantidad de registros(Antes) | Cantidad de registros(Después) |
|---|---|---|
| P+ | 146 | 186 |
| P- | 186 | 186 |

**Transformación de datos**
Para la transformación de los datos se consideró solo una tarea principal. Respecto al escalamiento de características numéricas, se utilizó la técnica de escalamiento estándar y se

aplicó en las 8 variables continuas como temperatura máxima, temperatura mínima, precipitación, latitud y longitud con el objetivo de normalizar los datos y así obtener una mejor precisión de predicción de los modelos.

### 2.3 Modelado.

**Selección de algoritmos.**
Se consideró los siguientes algoritmos: LRN, LDA, SVM, SGD, KNN, GNB, DTS, RFC, NNM, XGB a través de la librería scikit-learn. Por lo tanto, se hizo una evaluación con k-fold Cross Validation con 10 iteraciones de los datos de entrenamiento y se seleccionó al algoritmo que obtuvo una mayor precisión: RFC con 87.49%.

**Entrenamiento y prueba.**
Se realizó el entrenamiento con el 80% del conjunto de datos y el 20% para la prueba, usando el algoritmo RFC. Por otro lado, en la tabla 4, presenta las variables más importantes que aportan al modelo, siendo las mejores variables temp_max l_0, longitud y precip l_0.

**Table 4.** Nivel de importancia de las features

| N° | Features | Nivel de importancia |
|---|---|---|
| 1 | temp_max l_0 | 0.150712 |
| 2 | longitud | 0.143817 |
| 3 | precip l_0 | 0.131407 |
| 4 | latitud | 0.131009 |
| 5 | temp_max l_2 | 0.122904 |
| 6 | temp_min l_0 | 0.114546 |
| 7 | precip l_2 | 0.108171 |
| 8 | temp_min l_2 | 0.097435 |

### 2.4 Despliegue del modelo

**Desarrollo de la app web**
Se desarrolló un micro servicio usando el framework Flask de Python, con la finalidad de acceder al modelo de ML realizado en el paso anterior. Además, se generó una cuenta de acceso al API de weatherbit.io quien otorga datos climáticos (temperatura máxima, mínima y precipitación) pronosticados de 16 días a la fecha actual, esto se vinculó a la API REST.

Se generó una cuenta de pago para el acceso al API de angular-maps, esto permitió realizar las modificaciones de la presentación del google maps, resaltando de color rojo a las provincias con presencia de plagas y verde a las provincias sin presencia de plaga. Esto fue considerado dentro del API REST. Asimismo, se desarrolló la app cliente en Angular que consume los servicios de weatherbit.io [31], modelo predictivo de plagas y angular-maps mediante la API REST, mostrando los resultados de las plagas por provincia de la región San Martín en un google maps.

**Despliegue de la app web**
Para el despliegue del aplicativo de ML se utilizó herramientas como github [32] para versionar el producto y tener la disponibilidad de subir a cualquier servidor en la nube y también Heroku [33] esto es una plataforma que permite implementar y ejecutar una aplicación en la web ver figura 4.

## 3 Resultados

En la tabla 5, se presenta los algoritmos con mayor precisión que se obtuvo durante la fase del modelado, de los cuales el algoritmo Random Forest Classifier (RFC) fue el que obtuvo el mayor puntaje con 87.49% de accuracy..

**Table 5.** Resultado del entrenamiento de los algoritmos con mayor precisión.

| Algorithm | Accuracy |
|---|---|
| RFC | 87.49 |
| DTS | 85.59 |
| XGB | 85.22 |

En la tabla 6, se muestra los resultados del algoritmo seleccionado después de haber aplicado el balanceo y calibración del modelo, donde P+ representa la presencia de la plaga Burkholderia Glumae y P- representa la ausencia de la misma, donde se logró un 98.31% de accuracy en el entrenamiento y un 88% en la prueba respectivamente.

**Table 6.** Resultado de entrenamiento y prueba con el algoritmo RFC

| | TARGET | Precision | Recall | f1-score | Accuracy |
|---|---|---|---|---|---|
| Entrenamiento | P+ | 0.99 | 0.97 | 0.98 | 0.9831 |
| | P- | 0.97 | 0.99 | 0.98 | |
| Prueba | P+ | 0.84 | 0.91 | 0.87 | 0.88 |
| | P- | 0.92 | 0.85 | 0.89 | |

Asimismo, los resultados fueron muy alentadores, con una accuracy y recall muy eficientes, donde el modelo fue capaz de predecir a un 84% de precisión respecto a la presencia de la plaga Burkholderia Glumae y asegurar en un 91% que efectivamente sea la plaga. Además, se calculó el Área under the ROC Curve (AUC) de 65%. Este porcentaje indica que el modelo de ML tiene la probabilidad del 65% de distinguir entre las plagas negativas y positivas.

En la siguiente etapa, se logró cargar el modelo *.joblib a una arquitectura de desarrollo de software con lenguaje de programación Python del lado del servidor o backend y el framework Angular del lado del cliente o frontend.

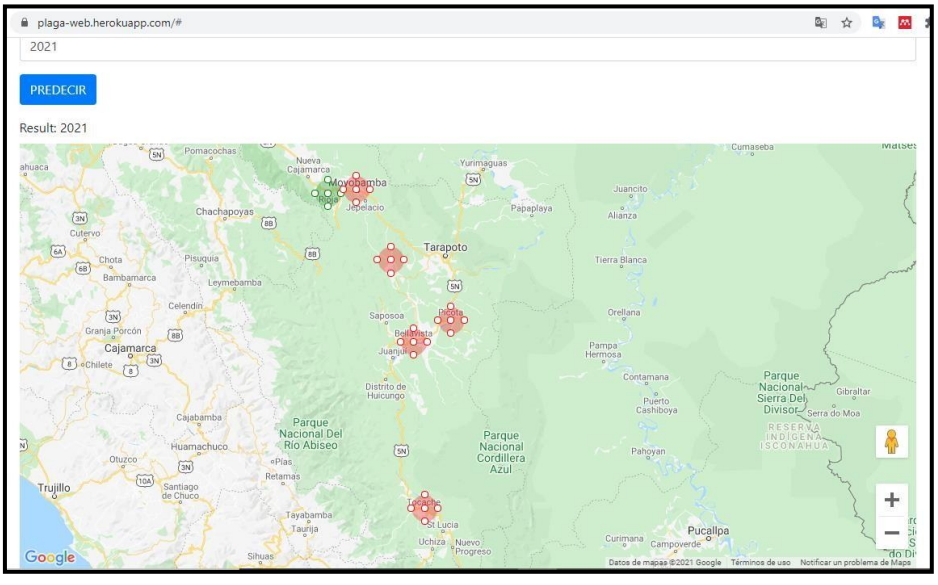


**Fig. 4.** Visualización de la predicción de la plaga Burkholderia Glumae, en las 6 provincias seleccionadas.

La fig. 4 muestra la interfaz del cliente presentando las 6 provincias (San Martín, Rioja, Moyobamba, Bellavista, Tocache y Picota), las marcadas de color rojo manifiestan presencia de plaga y la de color verde la ausencia de estas en los siguientes 16 días.

## 4 Conclusiones

Los datos compartidos por el SENASA, fueron de vital importancia debido a que contenían las pruebas fitosanitarias de un sembrío de arroz en particular, pero, carecían de información climática, en ese sentido, fue necesario la utilización de la técnica de IDW para completar la información. Los resultados obtenidos evidencian que el modelo presentó un 88% de accuracy y la capacidad de separabilidad entre la ausencia y presencia de la plaga a un 65%. Asimismo, la solución que se planteó frente a este problema fue tomar las medidas de prevención, respecto a la plaga Burkholderia Glumae en el cultivo de arroz en la región de san Martin. De igual modo, el estudio sugiere a la SENASA y a los investigadores en invertir en el sistema de vigilancia fitosanitaria, pues mientras más muestras existan en una zona determinada, se tendrá mejores resultados de predicción.

## References


[1] MINAGRI, “Plan Nacional De Cultivos Campaña Agrícola 2018-2019,” *Plan Nac. Cultiv.*, p. 323, 2017.

[2] SENASA, “San Martín: Monitoreo preventivo en cultivos de arroz - SENASA Contigo,” 2017. [Online]. Available: https://www.senasa.gob.pe/senasacontigo/san-martin-monitoreo-preventivo-en-cultivos-de-arroz/. [Accessed: 09-Jul-2018].

[3] K. Matthews, “6 Ways the Agricultural Industry Is Benefiting From Data Scientists,” *Towards Data Science*. [Online]. Available: https://towardsdatascience.com/6-ways-the-agricultural-industry-is-benefiting-from-data-scientists-b778d83f61db. [Accessed: 06-Jan-2021].

[4] D. Jiménez *et al.*, “A scalable scheme to implement data-driven agriculture for small-scale farmers,” *Glob. Food Sec.*, vol. 23, pp. 256–266, Dec. 2019.

[5] Z. Chen, H. Pan, C. Liu, and Z. Jiang, “Chapter 7 - Agricultural Remote Sensing and Data Science in China,” F. A. Batarseh and R. B. T.-F. D. S. Yang, Eds. Academic Press, 2018, pp. 95–108.

[6] W. S. Kang, S. S. Hong, Y. K. Han, K. R. Kim, S. G. Kim, and E. W. Park, “A web-based information system for plant disease forecast based on weather data at high spatial resolution,” *Plant Pathol. J.*, vol. 26, no. 1, pp. 37–48, 2010.

[7] R. Dhaya, “Flawless Identification of Fusarium Oxysporum in Tomato Plant Leaves by Machine Learning Algorithm,” *J. Innov. Image Process.*, vol. 2, no. 4, pp. 194–201, 2021.

[8] H. J. Deva Koresh, “Analysis of Soil Nutrients based on Potential Productivity Tests with Balanced Minerals for Maize-Chickpea Crop,” *J. Electron. Informatics*, vol. 3, no. 1, pp. 23–35, Mar. 2021.

[9] DRASAM, “Diagnóstico cadena arroz y maíz marzo 2016.” SAN MARTÍN, 2016.

[10] Organización de la Naciones Unidas para la Alimentación y la Agricultura(FAO), “Los mercados de productos básicos agrícolas: el comercio agrícola, el cambio climático y la seguridad alimentaria,” 2018.

[11] AGRONOTICIAS, “Revista Agronoticias,” *Revista*, 2018. [Online]. Available: https://agronoticias.pe/ciencia-e-innovacion/agricola/anublo-el-flagelo-del-arroz/. [Accessed: 04-Feb-2021].

[12] MINAGRI, “Generalidades del arroz,” 2015. [Online]. Available: http://minagri.gob.pe/portal/26-sector-agrario/arroz/217-generalidades-del-producto. [Accessed: 25-Jul-2018].

[13] MINAGRI, “Sistema Información de Cultivos - SISSIC,” 2021. [Online]. Available: http://sissic.minagri.gob.pe/sissic. [Accessed: 08-Jan-2021].

[14] A. Quesada Gonzales and F. García Santamaría, “Burkholderia glumae en el cultivo de arroz en Costa Rica,” vol. 25, no. 2, pp. 371–381, 2014.

[15] S. Gomathinayagam, M. Rekha, S. S. Murugan, and J. C. Jagessar, “The biological control of paddy disease brown spot (Bipolaris oryzae) by using Trichoderma viride in vitro condition,” *J. Biopestic.*, vol. 3, no. 1 SPEC.ISSUE, pp. 93–95, 2010.

[16] M. P. Kennedy *et al.*, “Burkholderia gladioli: Five year experience in a cystic fibrosis and lung transplantation center,” *J. Cyst. Fibros.*, vol. 6, no. 4, pp. 267–273, Jul. 2007.

[17] FAO, *Guia para identificar las limitaciones de campo en la produccion de arroz.* Food & Agriculture Organi, 2003.

[18] J. Carolina and G. Ruesta, “Modelo de pérdidas para determinar precipitación efectiva usando sistemas de información geográfica departamento de Ingeniería Civil,” 2004.

[19] N. S. Lam, “Spatial Interpolation,” *Int. Encycl. Hum. Geogr.*, pp. 369–376, 2009.

[20] G. Q. Tabios and J. D. Salas’, “Water resources bulletin a comparative analysis of techniques for spatial interpolation of precipitation,” 1985.

[21] B. Boehmke and B. Greenwell, *Hands-On Machine Learning with R*. 2019.

[22] H. I. Rhys, *Machine Learning with R, the tidyverse and mlr*. 2020.

[23] A. P. Yunus *et al.*, “Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan,” 2019.

[24] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” 2001.

[25] A. Cutler, D. R. Cutler, and J. R. Stevens, “Random Forests,” in *Ensemble Machine Learning*, Boston, MA: Springer US, 2012, pp. 157–175.

[26] N. Saboya, A. A. S. Peruvian, and O. L. Loaiza, “Predictive model based on machine learning for the detection of physically mistreated women in the Peruvian scope,” *ACM Int. Conf. Proceeding Ser.*, pp. 18–23, 2019.

[27] SENASA, “Servicio Nacional de Sanidad Agraria del Perú - SENASA | Gobierno del Perú,” 2021. [Online]. Available: https://www.gob.pe/senasa. [Accessed: 24-Aug-2021].

[28] SENAMHI, “Estaciones hidrometeorológicos,” 2021. [Online]. Available: https://www.senamhi.gob.pe/?&p=estaciones. [Accessed: 24-Aug-2021].

[29] SENAMHI, “SENAMHI - Perú,” 2021. [Online]. Available: https://www.senamhi.gob.pe/?p=pronostico-meteorologico. [Accessed: 24-Aug-2021].

[30] J. Franklin *et al.*, “Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones,” *Arch. Venez. Farmacol. y Ter.*, vol. 37, no. 5, pp. 587–595, 2018.

[31] Weatherbit.io, “Weatherbit | Weather API - Historical Weather API,” 2021. [Online]. Available: https://www.weatherbit.io/. [Accessed: 24-Aug-2021].

[32] RamirezJosue and J. Perez-Suarez, “plaga-web en github.” [Online]. Available: https://github.com/jhopes/plaga-web. [Accessed: 24-Aug-2021].

[33] HEROKU, “Cloud Application Platform | Heroku,” 2021. [Online]. Available: https://www.heroku.com/. [Accessed: 24-Aug-2021].

ANEXO

Springer

ICMCSI

*Certificate of Presentation*

This certificate is awarded to

**Joel Perez-Suarez**

for successfully presenting the paper entitled

**Prediction model of the Burkholderia Glumae pest in rice crops using Machine Learning and Spatial Interpolation**

at the
3rd International Conference on
Mobile Computing and Sustainable Informatics (ICMCSI 2022)
organized by Tribhuvan University, Nepal
on the 27th & 28th of January 2022.

Session Chair

Organizing Secretary
Dr. Jennifer S. Raj

Conference Chair
Prof. Dr. Subarna Shakya