Predicción del rendimiento de los cultivos

Integrantes: **Rut**
Francisco Gálvez , 19.620.398 - 2
Andrés Navarro , 20.787.149 - 4
Sebastián Ortega , 20.542.390 - 7

PARTE I : Cambios realizados durante la propuesta

A continuación se presentará en formato tabular los puntos que se han mejorado respecto al primer Entregable, así como el antes y después del proyecto.

Cabe señalar en primera instancia que el primero de nuestros entregables cambió bastante en relación a su objetivo y área. En nuestra primera entrega nuestro proyecto era “Predicción de tendencias de consumo en el Mercado Chileno”, este proyecto aunque llamaba fuertemente nuestra atención, la falta de datos que están dispuestos al público son pocas y poco concluyentes por lo que buscamos especificar el área de nuestro proyecto.
Luego de barajar opciones llegamos a una nueva propuesta, “Predicción del rendimiento de los cultivos”. La fuente de información inicial de este proyecto viene de Kaggle, con datos recolectados de dos fuentes principales. Por un lado la recolección de datos sobre pesticidas e historial de rendimiento de cultivos se obtuvieron de la Organización de las Naciones Unidas para la Alimentación y la Agricultura(FAO en inglés). Por otro lado, los datos de precipitaciones y temperatura promedio se obtienen de los datos del Banco Mundial. Toda esta información es de libre acceso.
En relación a la importancia de este proyecto, en base a documentos y lecturas que hemos investigado este proyecto se centra principalmente en los grandes productores, el objetivo es entregar una herramienta que los ayude a predecir, controlar y actuar en base a los resultados de este modelo para maximizar rendimientos.
En esta ocasión utilizaremos dos modelos que se recomiendan en las lecturas para predecir estos valores, los cuales son Random Forest y XGBoost ambos algoritmos de aprendizaje automatico utilizados para tareas de clasificacion y regresion, además de que ambos utilizan árboles de decisión para mejorar la precisión y el sobreajuste, cabe destacar que XGBoost es un modelo similar a Random Forest pero optimizado para aumentar la velocidad de ejecución.

Esperamos que las variables como pesticidas, precipitaciones y la temperatura promedio tengan una relación con el rendimiento de los cultivos , predecimos que habrá una alta correlación entre variables como temperatura promedio y el nivel de precipitaciones.
Junto con nuestro cambio de propuesta también debemos cambiar nuestros entregables anteriores, además de mejorarlos en relación a la entrega anterior por lo que nuestros nuevos entregables son los siguientes:

Debemos cumplir:

1.- Encontrar y estudiar variables que podrían tener una relación con la variable principal que en nuestro caso es el rendimiento de los cultivos

2.- Encontrar una base de datos con datos confiables y con la que podamos encontrar resultados concluyentes.

3.- Crear un modelo predictivo preciso que sirva como herramienta para tomar medidas preventivas como gestionar mejor los recursos o planificación de la siembra.

Esperamos cumplir:

1.- Identificar y analizar las variables más influyentes en el rendimiento de los cultivos, ya sea condiciones climáticas o intervención humana con pesticidas.

2.- Análisis de sensibilidad para determinar cuáles variables tienen un mayor efecto en el rendimiento de los cultivos, para que así los agricultores puedan enfocarse en otras variables para la producción.

3. Realizar un análisis exploratorio de los datos para identificar tendencias, patrones y relaciones entre los diferentes atributos de la base de datos.

Nos gustaría cumplir:

1.- Diseñar un sistema que no solo predice el rendimiento, sino que también ofrece recomendaciones sobre prácticas agrícolas como el uso adecuado de pesticidas basado en los resultados del modelo.
2.- En base a resultados entregados por el modelo, generar una propuesta de qué cultivos y sobre qué circunstancias es mejor invertir para obtener mejores rendimientos.

3.- Incorporar información socioeconómica y datos del mercado , para poder enriquecer nuestra dataset-proyecto y aumentar así la eficacia del modelo.
4. Investigar cómo diferentes escenarios climáticos futuros podrían afectar el rendimiento de los cultivos y cómo los agricultores pueden adaptarse a estos cambios.

PARTE II : Avances del proyecto

Clasificación de atributos

Según la clasificación propuesta por S.S Stevens los atributos de la base de datos a trabajar se pueden presentar de la siguiente manera.

Atributo	**Tipo de atributo**
Pais	Nominal
Producto	Nominal
Año	Ordinal
Promedio precipitaciones	Razón
Toneladas pesticida	Razón
Temperatura promedio	Intervalo
Rendimiento parcial	Razón

Con respecto a los “must deliverables” se han logrado según lo esperado. Se determinó que las variables más determinantes en relación al rendimiento de los cultivos son principalmente variables como la temperatura promedio, las precipitaciones promedio, así como una variable relevante en el ámbito agrícola que son los pesticidas debido a su influencia en el medio ambiente.
Sobre el origen de la base de datos, ésta fue encontrada en la página web “Kaggle”, donde reconociendo que no es tan fiable a primera vista como lo sería una organización gubernamental, tienen su fuente desde datos desde el Banco Mundial de Datos y la Organización de Alimentos y Agricultura de las Naciones Unidas. Sobre el último punto sobre crear un modelo de predicción a través de una regresión, esto se llevó a cabo a través de una técnica de Machine Learning siendo la de Random Forest Regressor, ya que según la naturaleza de los datos recolectados se decidió orientarse por esta técnica debido a su nivel de precisión y además que permite relacionar datos que no están relacionados linealmente.
En cuanto a las recomendaciones que se pueden generar, por un lado, la recomendación principal que se busca con este proyecto es generar propuestas de inversión con altos retornos en rendimiento en base a las circunstancias climatológicas, toneladas de pesticidas y el tipo de cultivo que estemos tomando en cuenta para proporcionar así una herramienta útil a los productores a la hora de planificar su producción. Por otro lado, cabe señalar que la mayoría de los atributos en la base de datos no son manejables por el humano, por ejemplo la temperatura promedio o los niveles de precipitaciones anuales. Sin embargo, se podría hacer un mayor control y por lo tanto recomendaciones sobre el uso de los pesticidas medidos en toneladas, ya que su uso desmedido afecta las napas subterráneas así como los ríos y canales fluviales en los alrededores de los cultivos lo que podría afectar de manera indirecta los rendimientos.

A continuación se mostrarán los modelos de regresión para la predicción de la variable continua rendimiento parcial de cultivos en base a nuestra data asociada, a través de una comparación entre ellos:

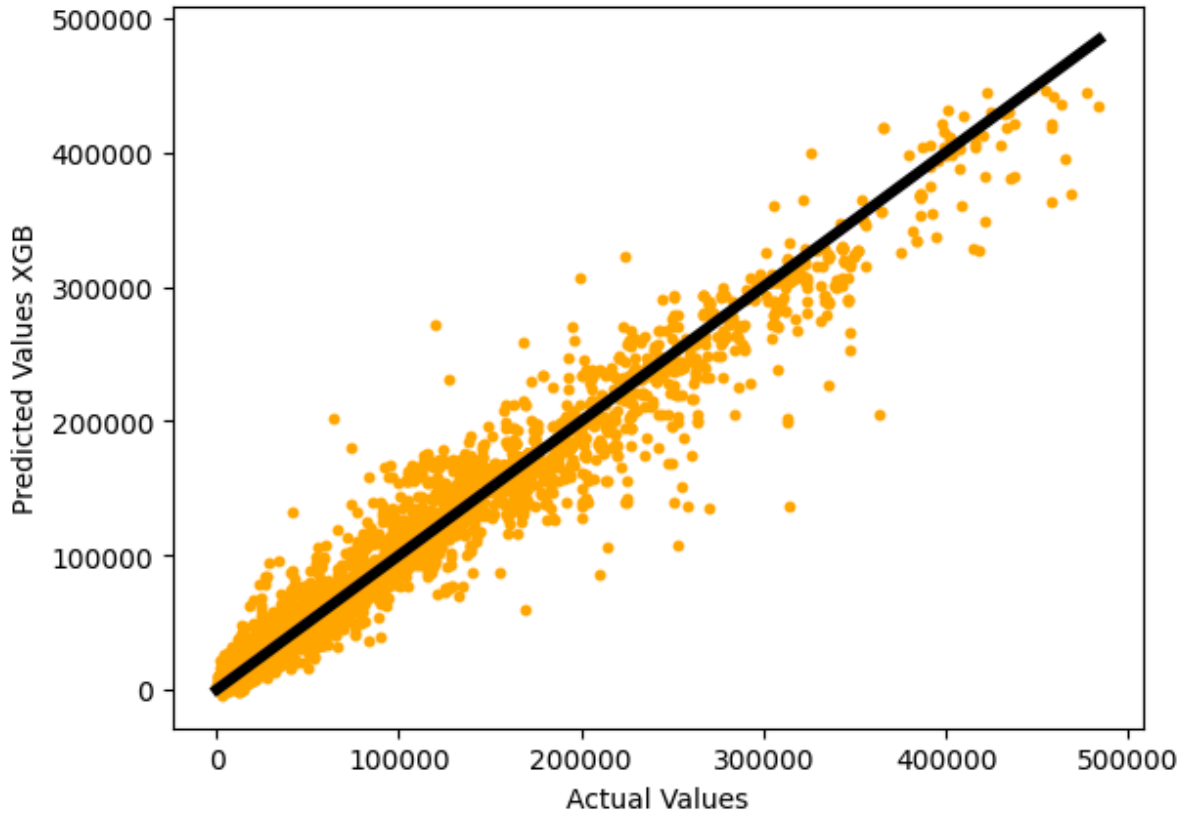
Diferenciación y análisis beneficios-costo de modelos utilizados
Random forest regressor: El beneficio principal es de facilidad de uso, donde no requiere tanto ajuste de hiperparámetros como otros métodos avanzados como

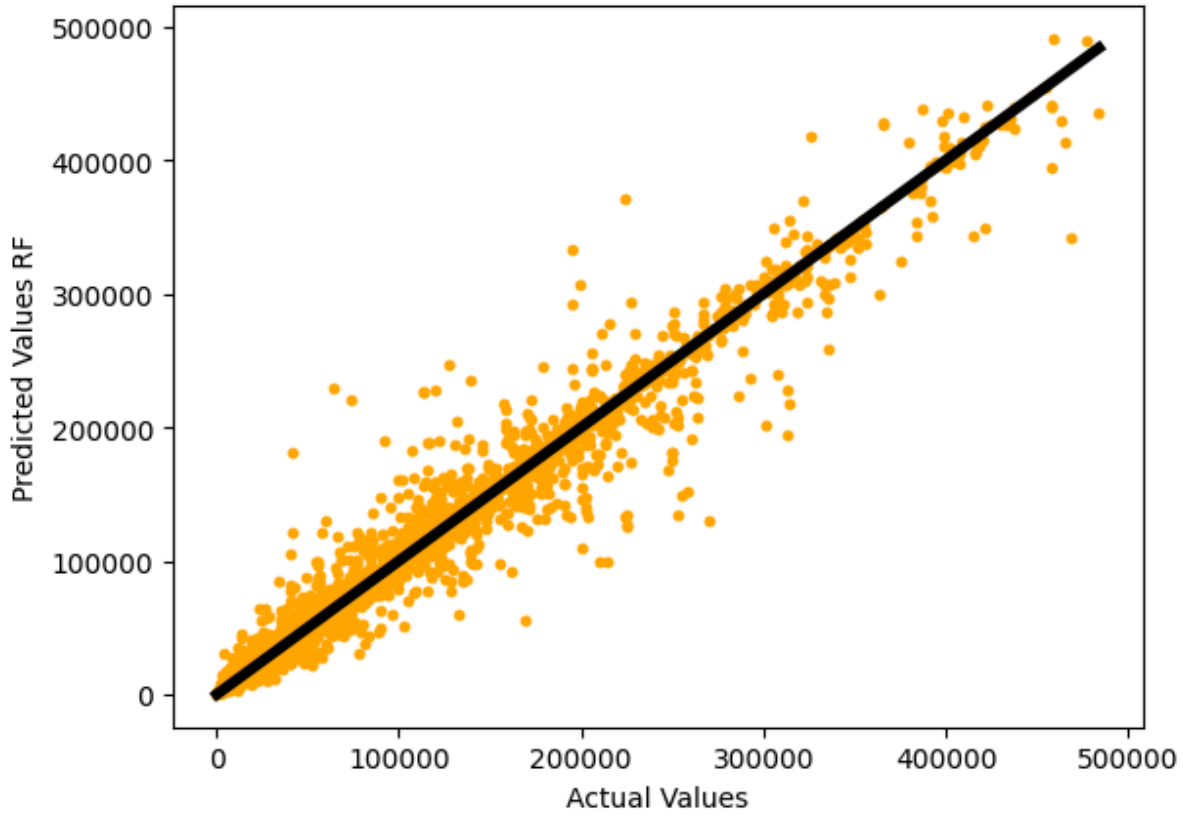
Gradient Boosting o XGBoost, lo que facilita su implementación en proyectos con menos experiencia en ciencia de datos.

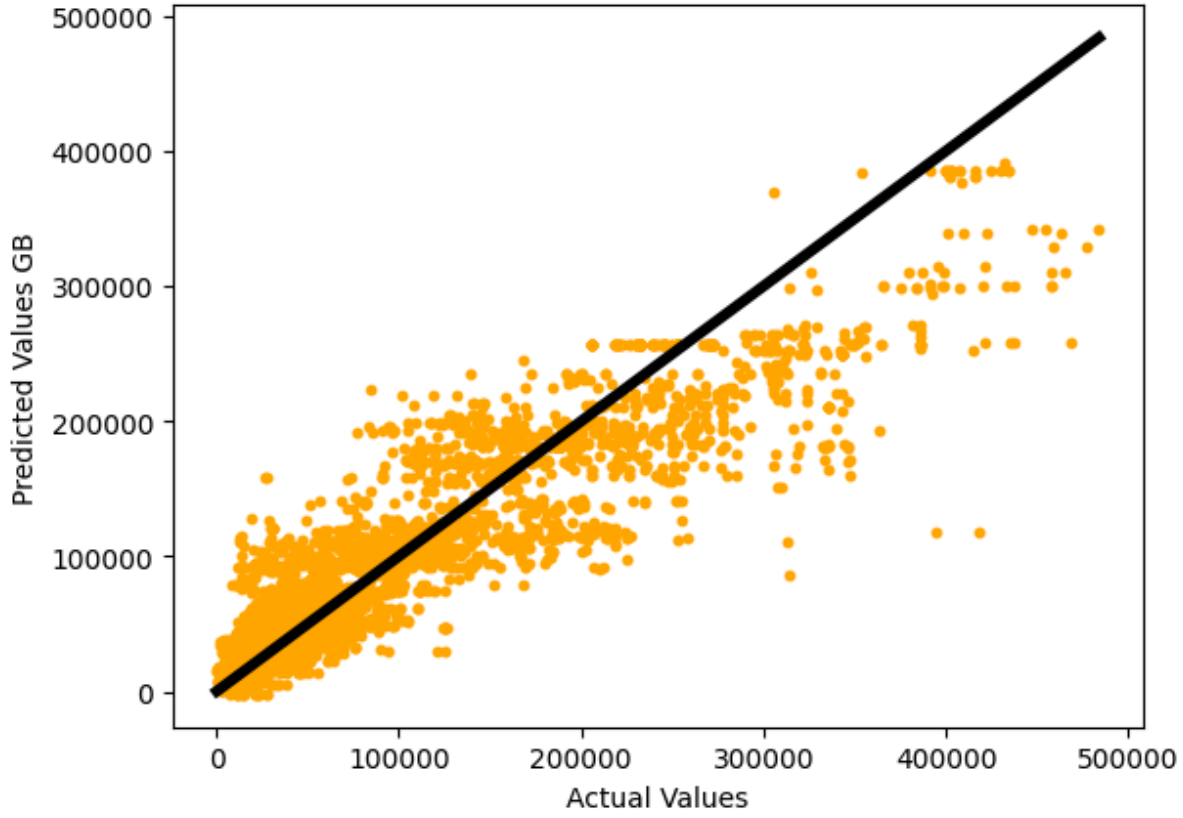
Gradient Boosting Regressor: Método de aprendizaje supervisado donde el objetivo es predecir un valor continuo.El modelo se basa en la técnica de boosting, que combina múltiples modelos débiles, donde cada árbol intenta corregir cada rama anterior. El gran costo que presenta es el sobreajuste donde si no configuramos bien los parámetros del modelo o usamos técnicas para evitar que se "sobreajuste", el modelo podría aprenderse demasiado los detalles específicos de los datos de entrenamiento. Esto es como si memorizará los ejemplos en lugar de entender las ideas generales, lo que haría que el modelo no funcione tan bien con datos nuevos. Por otro lado, su principal beneficio es su alta precisión; Gradient Boosting es conocido por su capacidad de capturar patrones complejos en los datos, lo que puede resultar en predicciones muy precisas.

XGBoost regressor: El principal beneficio de este modelo es su alta precisión y rendimiento: es muy eficaz para obtener modelos precisos debido a su capacidad de manejo de patrones complejos. Esto lo hace ideal para proyectos donde mejorar ligeramente la precisión puede tener un impacto significativo.
Por otro lado su principal desventaja es su riesgo de sobreajuste; Aunque XGBoost regressor ofrece técnicas de regularización, el riesgo de sobreajuste sigue presente, especialmente si no se ajustan bien los hiperparámetros.
Por otro lado su principal costo es que es menos interpretable, ya que aunque es más interpretable que otros métodos avanzados, como redes neuronales, sigue siendo un "modelo de caja negra" comparado con modelos lineales, lo que puede ser un inconveniente en aplicaciones donde se requiere una explicación clara de las decisiones del modelo.

Análisis de resultados obtenidos de modelos utilizados




Predicted Values RF
Actual Values
0
100000
200000
300000
400000
500000


Predicted Values GB
Actual Values
0
100000
200000
300000
400000
500000

Root Mean Squared Error (RMSE):
Gradient Boosting: 34042.56445850207
Random Forest (RF): 14,045.78
XGBoost Regressor (XGB): 16,564.78

El modelo de Random Forest Regressor tiene un RMSE más bajo, lo que indica que sus predicciones están más cerca de los valores reales en comparación con el modelo XGBoost Regressor (XGB) y Gradient Boosting. Es decir, el modelo RFR tiene un rendimiento más preciso en cuanto a la predicción de los datos.

Precision R cuadrado:

Gradient Boosting: 0.84023316748179
Random Forest Regressor (RFR): 0.9729
XGBoost Regressor (XGB): 0.9623

El valor de R^2 de 0.9729 en el modelo Random Forest Regressor(RFR) indica que este modelo explica aproximadamente el 97.29% de la varianza en los datos. Por otro lado, el modelo XGBoost Regressor (XGB), con un R^2 de 0.9623, explica el 96.23% de la varianza, lo que es ligeramente inferior y en relación al Gradient Boosting solo un 84,02%. Esto sugiere que Random Forest Regressor es un modelo más preciso en cuanto a la capacidad de explicar la variabilidad en los datos.

En cuanto a los modelos utilizados, se emplearon Random Forest regressor XGBoost Regressor y Gradient Boosting para predecir el rendimiento parcial por país. Con base en ambas métricas, el modelo de Random Forest regressor demostró un mejor desempeño que el modelo de XGBoost Regressor y Gradient Boosting, ya que presentó un RMSE más bajo y un R^2 más alto, lo cual indica que sus predicciones son más precisas y confiables.

Conclusión análisis

Después de analizar Gradient Boosting Regressor, XGBoost Regressor y Random Forest Regressor, hemos concluido que Random Forest Regressor es la

mejor opción para nuestro proyecto. Este modelo ofrece un equilibrio ideal entre precisión, estabilidad, y facilidad de implementación, lo cual es crucial cuando se trabaja con datos que pueden tener interacciones complejas o cuando se necesita una solución confiable sin un extenso ajuste de hiperparámetros.

Parte III

1-¿Creen que es necesario hacer cambios en los roles del equipo? (por ejemplo, en el spokeperson).

Consideramos que es importante evaluar la designación del speakperson alineándose con las fortalezas del equipo. Hemos notado que la comunicación no ha sido tan efectiva como esperábamos, además con un speakperson dinámico que se nutra de la retroalimentación hecha en cada entregable y consultas y/o reuniones con profesores proporciona mayores contribuciones en la comunicación de nuestro proyecto. Así el próximo speakperson se habrá enriqueciendo de las retroalimentaciones significando mejoras en las presentaciones
En cuanto a la designación de coordinador consideramos que una resignación vendrá en línea a intentar mejorar la eficiencia, ya que cambiar el coordinador a alguien con habilidades fuertes de gestión de proyectos puede resultar beneficioso para guiar el desarrollo del proyecto. Aceptando feedback del equipo y dispuesto a adaptarse a las necesidades del proyecto. Finalmente mencionar que el desarrollo fue exitoso en las entregas pero con una mayor experticia dentro de la gestión de proyectos, tal como fue desarrollándose en las entregas nos hubiera aminorado los tiempos de trabajo y de rediseño de proyecto.

2.-Reflexionen sobre cómo afectó la incertidumbre en un proyecto como este y cómo la enfrentaron, junto con qué decisiones tomaron y qué aprendieron de la experiencia.

Al principio nos sentíamos un poco perdidos, este fue nuestro primer acercamiento real a trabajar con bases de datos y por lo mismo no teníamos un conocimiento previo, por lo que nos vimos obligados a aprender y aplicar todo lo que se nos pedía en el proyecto, por suerte siempre tuvimos acceso a información y consultas sobre el proyecto por lo que pudimos guiarnos bien. Sentimos que si bien esta modalidad de presentarnos un proyecto tan complicado al principio sin entender realmente su magnitud y sus requerimientos puede ser un poco frustrante en primera instancia, nos presenta un desafío en el que debemos si o si usar lo aprendido en clases y además investigar y aplicar de manera individual, lo que refuerza el estudio y al menos en nuestro caso nos ha hecho sentirnos más preparados cuando se nos presentan estas problemáticas.

3.- ¿Cómo afectó la incertidumbre para el desarrollo del proyecto la elección de la base de datos? Reflexione al respecto de esto.

En nuestro caso fue lo que más nos complicó y limitó en este proyecto, al principio no sabíamos realmente el alcance de los datos que requerimos para un proyecto de este tipo, por lo que nos apegamos a una base creada por nosotros sin muchos objetos ni atributos, limitando nuestros resultados y con ello nuestra interpretación y reflexión. Si bien los datos eran provenientes de fuentes confiables como la ODEPA, INE o Dirección Meteorológica de Chile, no nos garantizó el éxito en hacer un modelo predictivo eficaz, por lo que tuvimos que prescindir de tal idea en su totalidad. Al final nuestro proyecto se mantuvo en la línea agrícola, aventurándonos por la predicción del rendimiento de los cultivos a nivel mundial, desde una base de datos que encontramos posteriormente y que nos permitió hacer los procesos de manera correcta, cómo encontrar correlaciones fuertes, poder aplicar de manera correcta los modelos de regresión y obtener resultados que si tuvieran una interpretación y que pudieran ser utilizados para generar una propuesta.

Como reflexión final, podemos darnos cuenta que para abordar efectivamente un problema de Data Science que involucre eventualmente el uso de herramientas de Machine Learning para finalmente ayudar a la toma de decisiones, implica un proceso de altos y bajos, pero que con el trabajo en equipo y en la búsqueda de nuevas ideas surgen propuestas que direccionan el trabajo final sacando a flote. Se toma en cuenta la importancia del preprocesamiento de la data y limpieza antes de realizar cualquier análisis descriptivo o inferencial más hoy en día con la abundante información y datasets en internet, como ocurre con la página web Kaggle y el repositorio de bases de datos. Finalmente concluimos en esta parte que aunque valga la redundancia lo más importante son los datos que se utilizaran, pues son el pilar fundamental para hacer este tipo de trabajo.

Bibliografia

Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha and V. G. R. R. Devi, "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 736-741, doi: 10.1109/ICCES48766.2020.9137868. keywords: {Temperature distribution;Machine learning algorithms;Input variables;Supervised learning;Crops;Production;Humidity;Supervised Learning;Naïve Bayes Algorithm;Regression;Decision Trees;Plots etc},

R. Medar, V. S. Rajpurohit and S. Shweta, "Crop Yield Prediction using Machine Learning Techniques," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033611. keywords: {Agriculture;Prediction algorithms;Machine learning;Clustering algorithms;Classification algorithms;Testing;Production;Indian Agriculture;Machine Learning Techniques;Crop selection method},

Abbas F, Afzaal H, Farooque AA, Tang S. Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. *Agronomy*. 2020; 10(7):1046. https://doi.org/10.3390/agronomy10071046