UNIVERSIDAD TECNOLÓGICA METROPOLITANA
del Estado de Chile

FACULTAD DE INGENIERÍA - ESCUELA DE INFORMÁTICA

INTELIGENCIA ARTIFICIAL (AI)

AI 2025-II. LAB 01

INTELIGENCIA ARTIFICIAL

EFEB6114 – 21041 / 21030 / 21049

INGENIERÍA CIVIL EN COMPUTACIÓN M / INFORMÁTICA
INGENIERÍA INFORMÁTICA

Departamento de Informática y Computación
Facultad de Ingeniería

Dr. Oscar Magna V.

LAB 01

Fecha de entrega: martes 23 de septiembre de 2025 (12h.)

ACTIVIDAD
Proyecto de Laboratorio N°1

TABLA DE CONTENIDO

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

PROYECTO

Predicción de Rendimiento de Cultivos en Agricultura de Precisión a Partir de Condiciones del Suelo y Clima

El proyecto LAB01 combina un contexto real de alto impacto - la seguridad alimentaria y la sostenibilidad agrícola - con el uso de modelos fundamentales de IA como el MLP, se fomenta tanto el aprendizaje técnico como la aplicación crítica del conocimiento. Los grupos de proyectos tendrán la oportunidad de trabajar con datos reales, enfrentar desafíos de preprocesamiento y modelado, y entregar un producto técnico con potencial aplicado en el mundo agrícola moderno.

1. CONTEXTO

La agricultura global enfrenta desafíos sin precedentes debido al cambio climático, la degradación de los recursos naturales, la creciente demanda alimentaria y la necesidad de operar con mayor sostenibilidad. Según estimaciones de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), se requiere un aumento del 70 % en la producción agrícola para el año 2050 para satisfacer las necesidades de una población mundial proyectada en más de 9 000 millones de personas. En este escenario, la agricultura de precisión se posiciona como una estrategia clave para mejorar la eficiencia productiva, optimizar el uso de insumos (agua, fertilizantes, energía) y reducir el impacto ambiental.

Uno de los pilares fundamentales de la agricultura de precisión es la toma de decisiones basada en datos, donde el análisis predictivo juega un rol central. La capacidad de anticipar el rendimiento de un cultivo permite a los agricultores planificar labores de campo, ajustar prácticas de manejo, gestionar riesgos climáticos y mejorar la rentabilidad. Aunque existen modelos agronómicos avanzados basados en simulaciones físicas (como DSSAT o APSIM), su complejidad técnica, alto costo computacional y necesidad de experticia limitan su adopción en pequeñas y medianas explotaciones.

En este contexto, los modelos de Inteligencia Artificial (IA) emergen como una alternativa accesible y escalable. En particular, el Perceptrón Multicapa (MLP), como modelo de red neuronal artificial feedforward, es altamente efectivo para modelar relaciones no lineales entre múltiples variables de entrada y una salida continua, como el rendimiento del cultivo. Aunque no está diseñado para procesar secuencias temporales largas, el MLP puede trabajar eficazmente con datos agregados por ciclo de cultivo (promedios, acumulados, índices), lo que lo hace adecuado para aplicaciones prácticas en entornos reales.

Este proyecto se enmarca en la necesidad de desarrollar soluciones tecnológicas basadas en IA que, sin requerir infraestructura costosa ni modelos complejos, permitan a los agricultores y técnicos tomar decisiones más informadas. Al utilizar arquitecturas fundamentales como el MLP, se busca no solo obtener resultados predictivos confiables, sino también fomentar la comprensión profunda de los fundamentos del aprendizaje automático en contextos aplicados.

2. Alcance

El proyecto se enfoca en la predicción del rendimiento de cultivos anuales de alto impacto (como maíz, trigo o soja) utilizando datos históricos provenientes de campos instrumentados o bases de datos públicas. El modelo será desarrollado bajo un enfoque de regresión supervisada, implementado exclusivamente con una arquitectura MLP, sin el uso de modelos secuenciales (RNN, LSTM) ni convolucionales (CNN).

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

El sistema incluirá un pipeline completo desde la adquisición de datos hasta la evaluación del modelo, pero no contemplará despliegue en tiempo real, interfaces gráficas avanzadas ni integración con sensores en campo. El trabajo se realizará sobre un único tipo de cultivo (a elección del equipo), utilizando datos agregados por ciclo de cultivo (por ejemplo, promedios mensuales o acumulados durante la temporada).

3. Objetivo General

Diseñar e implementar un modelo de red neuronal artificial tipo MLP (Multilayer Perceptron) capaz de predecir el rendimiento de cultivos (en toneladas por hectárea) a partir de variables edáficas (suelo), climáticas y de manejo agronómico, con el fin de apoyar la toma de decisiones en sistemas de agricultura de precisión.

4. Objetivos Específicos

1. Recopilar y preprocesar un conjunto de datos estructurado que incluya variables del suelo (humedad, pH, contenido de nitrógeno), clima (temperatura, precipitación, horas de sol) y prácticas de manejo (fecha de siembra, densidad, tipo de fertilización).
2. Realizar un análisis exploratorio de datos (EDA) para identificar correlaciones, valores atípicos y patrones significativos entre las variables de entrada y el rendimiento del cultivo.
3. Diseñar, entrenar y optimizar un modelo MLP mediante ajuste de hiperparámetros *(número de capas, neuronas, funciones de activación, tasa de aprendizaje)* y validación cruzada.
4. Evaluar el desempeño del modelo utilizando métricas de regresión *(RMSE, MAE, R^2, RMSE, MAPE, etc.)* y compararlo con modelos baseline tradicionales [*(Regresión lineal, Regresión con Regularización, Árbol de Decisión, Random Forest Regressor, Support Vector Regressor (SVR), K-Nearest Neighbors Regressor (KNN), árboles de decisión)*] y Modelos Modernos [*Regresión con Regularización (Ridge y Lasso), Árbol de Decisión (Decision Tree Regressor), Random Forest Regressor, Support Vector Regressor (SVR), K-Nearest Neighbors Regressor (KNN), Gradient Boosting (XGBoost), LightGBM, CatBoost, Extra Trees Regressor, HistGradientBoosting Regressor, Redes Neuronales Tabulares (MLP + mejoras), TabNet, AutoML (H2O, AutoGluon, TPOT)*].
5. Analizar la importancia relativa de las variables de entrada mediante técnicas como permutación de características o análisis de sensibilidad, con el fin de aportar interpretabilidad al modelo.

5. Hipótesis

El proyecto se sustenta en las siguientes tres hipótesis que permiten evaluar tanto el desempeño técnico del modelo MLP como su aplicabilidad práctica en el contexto agronómico:

1. Hipótesis principal (H_1):

Un modelo MLP entrenado con datos integrados de suelo, clima y manejo agronómico puede predecir el rendimiento de cultivos con un error absoluto medio (MAE) inferior al 10 % del rendimiento promedio observado en el conjunto de datos, superando significativamente el desempeño de modelos lineales simples.

2. Hipótesis secundaria (H_2):

Las variables climáticas (especialmente precipitación acumulada y temperatura media) y las relacionadas con el contenido de nitrógeno en el suelo tienen un peso significativamente mayor en la predicción del rendimiento que otras variables, según lo evidenciado por técnicas de análisis de importancia de características.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

3. Hipótesis de interpretabilidad (H_3):

A pesar de su naturaleza de "caja negra", un modelo MLP bien entrenado y analizado mediante técnicas como la permutación de características o el análisis de sensibilidad puede ofrecer insights interpretables y alineados con conocimientos agronómicos establecidos, permitiendo su uso como herramienta de apoyo a la toma de decisiones.

6. Interrogantes a Resolver

1. ¿Qué combinación de variables del suelo, clima y manejo agronómico tiene mayor impacto en la predicción del rendimiento?
2. ¿Cuál es la arquitectura óptima del MLP (número de capas, neuronas, función de activación) para este problema de regresión?
3. ¿Cómo afecta la calidad y completitud de los datos (valores faltantes, ruido, desbalance) al desempeño del modelo?
4. ¿Es posible obtener una predicción precisa del rendimiento sin utilizar datos de series temporales largas o imágenes satelitales?
5. ¿Qué nivel de interpretabilidad puede alcanzarse en un modelo MLP aplicado a un dominio agronómico crítico?

7. Consideraciones generales para el diseño

a) Pipeline propuesto

El desarrollo del proyecto seguirá un pipeline estructurado compuesto por las siguientes etapas:

1. Adquisición de datos: Descarga o generación de un conjunto de datos estructurado con variables agronómicas y rendimiento.
2. Preprocesamiento: Limpieza *(imputación de datos faltantes, detección de outliers)*, normalización (Min-Max o StandardScaler) y construcción de características derivadas (*acumulados, promedios, índices simples).*
3. Análisis exploratorio (EDA): Visualización de distribuciones, correlaciones y tendencias mediante gráficos *(histogramas, mapas de calor, boxplots).*
4. División de datos: Separación en conjuntos de entrenamiento (70 %), validación (15 %) y prueba (15 %).
5. Diseño del MLP: Implementación del modelo usando bibliotecas como TensorFlow/Keras o scikit-learn.
6. Entrenamiento y validación: Ajuste con validación cruzada, uso de early stopping y técnicas de regularización.
7. Evaluación: Cálculo de métricas *(RMSE, MAE, R^2, RMSE, MAPE, etc.)* y comparación con modelos baseline *(tradicionales y modernos).*
8. Interpretabilidad: Análisis de importancia de características mediante permutación o gradiente.
9. Documentación: Redacción del informe técnico, código comentado y presentación de resultados.

b) Conjunto de Datos

El modelo se entrenará con un conjunto de datos estructurado que incluya, por cada observación (parcela + ciclo de cultivo):

1. Variables de entrada:
    - Suelo: pH, humedad, contenido de nitrógeno (N), fósforo (P), potasio (K), materia orgánica.
    - Clima: precipitación acumulada, temperatura media, horas de sol.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

- Manejo: tipo de cultivo, densidad de siembra, fecha de siembra, uso de fertilizantes.

2. Variable objetivo: Rendimiento (toneladas por hectárea).

c) Tareas Generales del Proyecto

Los grupos de proyectos deberán realizar, al menos, las siguientes tareas:

- Revisión bibliográfica *(considerar las bases de datos propuesta en el Marco de Trabajo- Proyectos de Titulación y de las Cátedras de Gestión de Proyecto Informáticos y AI)* sobre agricultura de precisión y modelos de predicción de rendimiento.
- Selección y preparación del dataset.
- Implementación del pipeline de preprocesamiento y análisis exploratorio.
- Diseño, entrenamiento y validación del modelo MLP.
- Evaluación comparativa con modelos baseline.
- Análisis de resultados e interpretación agronómica.
- Redacción del informe final y presentación de resultados.

d) Consideraciones especiales

- No se permitirá el uso de modelos secuenciales (RNN, LSTM) ni convolucionales (CNN).
- El enfoque debe ser supervisado y de regresión.
- Se debe priorizar la interpretabilidad del modelo, evitando arquitecturas excesivamente complejas.
- Se debe documentar el tratamiento de datos faltantes y el impacto de la normalización.
- Se recomienda el uso de técnicas de regularización (dropout, L2) para evitar sobreajuste.
- El modelo debe ser reproducible y el código debe estar bien estructurado y comentado. El modelo final debe ser producto de estrategias y acciones de mejora respecto al modelo seleccionado (este modelo es seleccionado a partir del análisis, comparación y evaluación del conjunto de modelos iniciales y preseleccionados).

e) Bases de Datos (Referencias iniciales)

Fuentes de datos públicas recomendadas para el desarrollo del proyecto:

1. USDA NASS Quick Stats
   - Base de datos oficial del Departamento de Agricultura de EE.UU. con estadísticas de rendimiento, área sembrada y prácticas agrícolas.
2. Kaggle – Crop Yield Prediction
   - Datasets como "USA Corn Yield" o "Crop Yield in india" con variables climáticas y de manejo.
3. FAOSTAT
   - Estadísticas agrícolas globales sobre producción, rendimiento y uso de tierras por país.
4. The Open Ag Data Alliance (OADA)
   - Plataforma de código abierto para compartir datos agrícolas de forma segura e interoperable.
5. Agrimetrics (UK)
   - Plataforma de datos integrados (clima, suelo, mercado) con acceso parcial gratuito para investigación.
6. Climate Data Online (NOAA)
   - Datos climáticos históricos (temperatura, precipitación) por ubicación geográfica.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

8. Tareas Generales del Proyecto

El desarrollo del proyecto se organiza en seis fases clave, cada una compuesta por tareas específicas que deben ser ejecutadas de forma secuencial y colaborativa. El proyecto tiene una duración estimada de 4 semanas, por lo que se recomienda seguir un cronograma semanal.

• Fase 1: Investigación Inicial y Planificación del Proyecto

TAREA	DESCRIPCIÓN	ENTREGABLE
1.1 Revisión bibliográfica	Investigar trabajos previos sobre predicción de rendimiento agrícola, modelos de IA en agricultura de precisión y uso de MLP en datos tabulares.	Documento de revisión con al menos 5 fuentes científicas, seleccionadas (analizadas y evaluadas) del conjunto potencial de relevamiento bibliográfico.
1.2 Definición del alcance	Acordar el cultivo objetivo (maíz, trigo, soja, etc.), región geográfica de enfoque y variables clave a incluir.	Acta de reunión con alcance definido.
1.3 Diseño del cronograma	Elaborar un cronograma de trabajo semanal con hitos, responsables y entregables.	Gantt o tabla de actividades por semana.
1.4 Selección de herramientas	Elegir lenguaje (Python), librerías (scikit-learn, TensorFlow, pandas, matplotlib), entorno (Jupyter, Colab, VS Code) y formato	Fundamentos técnicos considerados3

• Fase 2: Adquisición y Preparación de Datos

TAREA	DESCRIPCIÓN	ENTREGABLE
2.1 Búsqueda de datasets	Identificar y descargar datos de fuentes públicas (USDA, FAOSTAT, Kaggle, NOAA).	Lista de datasets evaluados con URL y descripción.
2.2 Integración de datos	Unificar datos de suelo, clima y rendimiento en un único dataset estructurado (CSV/Parquet).	Dataset unificado listo para análisis.
2.3 Limpieza de datos	Tratar valores faltantes (imputación), detectar y gestionar outliers, eliminar duplicados.	Informe de limpieza con justificación de decisiones.
2.4 Ingeniería de características	Crear nuevas variables útiles: acumulados de lluvia, grados-día, índices de fertilidad, categorías de clima.	Conjunto de features derivadas documentadas.

• Fase 3: Análisis Exploratorio de Datos (EDA)

TAREA	DESCRIPCIÓN	ENTREGABLE
3.1 Estadísticas descriptivas	Calcular media, desviación, percentiles, valores mínimos/máximos por variable.	Tabla de estadísticas.
3.2 Visualización de distribuciones	Graficar histogramas, boxplots y densidades para identificar sesgos o asimetrías.	Panel de gráficos (matplotlib/seaborn).
3.3 Matriz de correlación	Analizar correlaciones entre variables y con el rendimiento (uso de mapa de calor).	Mapa de calor con interpretación.
3.4 Análisis por temporada/clima	Segmentar datos por año, estación o región para detectar patrones.	Gráficos comparativos por grupo.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

- Fase 4: Diseño, Entrenamiento y Validación de Modelos

TAREA	DESCRIPCIÓN	ENTREGABLE
4.1 División de datos	Separar el dataset en entrenamiento (70%), validación (15%) y prueba (15%).	Conjuntos guardados y documentados.
4.2 Implementación de modelos baseline	Entrenar los 6 modelos clásicos: Regresión Lineal, Ridge, Lasso, Árbol, RF, SVR, KNN.	Código modular y resultados por modelo.
4.3 Implementación de modelos modernos	Entrenar al menos 3 modelos avanzados: XGBoost, LightGBM, CatBoost (y opcionalmente TabNet o AutoML).	Código y métricas de modelos modernos.
4.4 Diseño y entrenamiento del MLP	Definir arquitectura (capas, neuronas, activaciones), entrenar con validación cruzada y early stopping.	Modelo MLP guardado y documentado.
4.5 Ajuste de hiperparámetros	Usar Grid Search o Random Search para optimizar hiperparámetros en al menos 3 modelos (incluyendo el MLP).	Tabla de mejores hiperparámetros.
4.6 Evaluación comparativa	Calcular MAE, RMSE, R^2 y MAPE para todos los modelos en el conjunto de prueba.	Tabla comparativa completa.

- Fase 5: Análisis de Resultados e Interpretabilidad

TAREA	DESCRIPCIÓN	ENTREGABLE
5.1 Comparación de desempeño	Identificar el modelo con mejor equilibrio entre precisión, eficiencia e interpretabilidad.	Gráfico de barras de métricas.
5.2 Análisis de residuos	Graficar errores vs. valores predichos para detectar sesgos sistemáticos.	Scatter plot de residuos.
5.3 Importancia de características	Usar permutation importance o SHAP para identificar variables más influyentes en el MLP y modelos como XGBoost (o determinarlo con otras técnicas).	Gráfico de importancia de features.
5.4 Validación de hipótesis	Evaluar si se aceptan o rechazan las tres hipótesis del proyecto.	Sección de conclusiones sobre hipótesis.
5.5 Discusión técnica	Analizar por qué ciertos modelos superan a otros (ej: ¿por qué XGBoost > MLP?).	Párrafo de discusión en informe.

- Fase 6: Análisis de Resultados e Interpretabilidad

TAREA	DESCRIPCIÓN	ENTREGABLE
6.1 Redacción del informe técnico	Elaborar un documento completo con todas las secciones: contexto, objetivos, metodología, resultados, conclusiones.	Informe final en PDF (8–12 páginas).
6.2 Documentación del código	Comentar el código, incluir unREADME.md, y estructurarlo en módulos (datos, modelos, evaluación).	Repositorio GitHub o carpeta organizada.
6.3 Preparación de presentación	Crear una presentación (PowerPoint o Google Slides) con máximo 10 diapositivas.	Presentación visual clara.
6.4 Prueba de demostración	Ejecutar el modelo con un ejemplo de predicción nueva (ej: un campo con datos simulados).	Demo funcional (script o notebook).
6.5 Entrega final	Subir todos los archivos (informe, código, presentación, dataset reducido) al sistema de entrega.	Carpeta comprimida.zipo enlace a repositorio.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

- Herramientas Recomendadas

TAREA	DESCRIPCIÓN	ENTREGABLE
1. Programación	Python (Jupyter Notebook, Google Colab, VS Code)	Programación
2. Gestión de datos	pandas, NumPy	Gestión de datos
3. Visualización	matplotlib, seaborn, plotly	Visualización
4. Modelos baseline tra		
5. Modelos baseline tradicionales (clásicos)	Modelos en scikit-learn (clasificación, regresión y Clustering). Considerar: Regresión lineal, Regresión con Regularización, Árbol de Decisión, Random Forest Regressor, Support Vector Regressor (SVR), K-Nearest Neighbors Regressor (KNN), árboles de decisión	Modelos clásicos
6. Modelos modernos	Considerar: Regresión con Regularización (Ridge y Lasso), Árbol de Decisión (Decision Tree Regressor), Random Forest Regressor, Support Vector Regressor (SVR), K-Nearest Neighbors Regressor (KNN), Gradient Boosting (XGBoost), LightGBM, CatBoost, Extra Trees Regressor, HistGradientBoosting Regressor, Redes Neuronales Tabulares (MLP + mejoras), TabNet, AutoML (H2O, AutoGluon, TPOT)	Modelos modernos
7. AutoML	H2O.ai, AutoGluon, TPOT	AutoML
8. Documentación	Markdown, LaTeX (opcional), Google Docs	Documentación
9. Control de versiones	GitHub/GitLab	Control de versiones

- Conclusión

Una vez evaluado todos los modelos, los grupos deberán:

1. Definir objetivos de mejora, determinando estrategias de mejora y acciones, las cuales deberán implementarse para identificar y seleccionar el modelo con el mejor equilibrio entre precisión, interpretabilidad y eficiencia
2. Justificar por qué el MLP supera (o no) a los modelos.
3. Discutir si el aumento de complejidad del MLP está justificado por la mejora en el desempeño.
4. Responder las hipótesis
5. Responder de manera fundamentada las interrogantes.
6. Definir un conjunto de recomendaciones

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

ANEXO:

N°1: Bases de Datos *(Referencias iniciales)*

1. Detección de Fraudes en Transacciones Financieras

- *Credit Card Fraud Detection*
  - **Descripción**: Contiene transacciones realizadas con tarjetas de crédito por clientes europeos en septiembre de 2013. Incluye 284,807 transacciones, de las cuales 492 son fraudulentas.
  - **Características**: Datos numéricos transformados mediante PCA, altamente desbalanceado.
  - **Enlace**: https://www.kaggle.com/mlg-ulb/creditcardfraud
- Fraudulent Transactions Dataset
  - **Descripción**: Conjunto más reciente que simula una variedad de tipos de fraudes en comercio electrónico.
  - **Enlace**: https://www.kaggle.com/ealaxi/paysim; https://datarade.ai/search/products/financial-fraud-detection-dataset
  - Synthetic Financial Datasets for Fraud Detection: Diseñado para evaluar algoritmos de detección de fraudes. https://datarade.ai/search/products/financial-fraud-detection-dataset
  - ***IEEE-CIS Fraud Detection Dataset:***
  - Contiene datos reales y sintéticos sobre transacciones financieras, ampliamente utilizado en investigaciones sobre detección de fraudes. https://datarade.ai/search/products/financial-fraud-detection-dataset

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

N°1: Guía de Auto documentación de Notebooks para Proyecto

Esta guía establece los estándares para la auto documentación del notebook de Jupyter/Google Colab que se desarrollará en el proyecto de "Predicción de Rendimiento de Cultivos en Agricultura de Precisión". Esta documentación facilitará la comprensión del código, el seguimiento del proceso de desarrollo y la evaluación del proyecto.

1. Estructura General del Notebook

El notebook debe seguir una estructura clara y coherente que facilite la navegación y comprensión:

1.1 Sección de Encabezado

Al inicio de cada notebook debe incluirse:

- [Título del Notebook] - Sistema de Seguridad Biométrico
- Grupo [Número de Grupo]
- Nombre Estudiante 1] - [Rol en el proyecto]
- Nombre Estudiante 2] - [Rol en el proyecto]
- Nombre Estudiante 3] - [Rol en el proyecto]
- Nombre Estudiante 4] - [Rol en el proyecto]

- Fecha: [Fecha de última actualización]
- Versión: [Número de versión]

- Objetivo del Notebook
- Descripción breve del propósito y objetivos específicos de este notebook

- Tabla de Contenidos
  1. [Primera sección principal](#1)
  2. [Segunda sección principal](#2)
  ...

1.2 Organización por Secciones

Dividir el notebook en secciones claras utilizando encabezados markdown:

- Nivel 1 (#) para el título principal
- Nivel 2 (##) para secciones principales
- Nivel 3 (###) para subsecciones
- Nivel 4 (####) para componentes específicos

1.3 Tabla de Contenidos

Incluir una tabla de contenidos con hipervínculos a las diferentes secciones del notebook para facilitar la navegación.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

2. Elementos Visuales

2.1 Gráficos y Visualizaciones

- Cada visualización debe incluir:
  - Título descriptivo
  - Etiquetas claras en los ejes
  - Leyenda cuando corresponda
  - Interpretación textual de lo que muestra

2.2 Tablas de Resultados

- Formato claro con encabezados descriptivos
- Destacar valores importantes
- Incluir unidades de medida cuando corresponda

2.3 Diagramas Conceptuales

- Incluir diagramas para explicar conceptos complejos
- Utilizar herramientas como Mermaid para diagramas en el propio notebook

3. Documentación del Proceso de Desarrollo

3.1 Gestión de datos

- **Carga de datos**: Verifica que los datos (e.g., CSV, bases de datos) se carguen correctamente, ya sea desde Google Drive, una URL o un dataset subido a Colab.
- **Exploración inicial**: Incluye un análisis exploratorio (EDA) con visualizaciones (histogramas, correlaciones, boxplots) para entender las variables de suelo y clima.
- **Limpieza de datos**:
  - Maneja valores faltantes (imputación, eliminación o justificación).
  - Detecta y trata outliers (e.g., valores extremos de pH o precipitación).
  - Verifica tipos de datos (e.g., numéricos para pH, categóricos para tipo de suelo).
- **Preprocesamiento**:
  - Normaliza o estandariza variables numéricas (e.g., usando StandardScaler de scikit-learn).
  - Codifica variables categóricas (e.g., tipo de cultivo o suelo con *one-hot encoding*).
  - Divide los datos en conjuntos de entrenamiento, validación y prueba (e.g., 70-20-10).
- **Reproducibilidad**: Fija una semilla (*random seed*) para cualquier proceso aleatorio (e.g., random_state=42 en scikit-learn).

3.2. Modelado

- **Selección de modelo(s)**:
  - Justificar la elección del modelo (e.g., regresión lineal, árboles de decisión, random forest, XGBoost, etc., tanto tradicionales como modernos) según el problema.
  - Obtener, con fundamentos, un conjunto potencial de modelos y sleecionar, al menos, 2-3 modelos como candidatos para evaluar cuál es más adecuado.
- **Entrenamiento**:
  - Entrena el modelo con los datos preprocesados.
  - Usa validación cruzada (e.g., cross_val_score) para evaluar la robustez.
- **Optimización de hiperparámetros**:
  - Aplica búsqueda de hiperparámetros (e.g., GridSearchCV o RandomizedSearchCV) para mejorar el rendimiento.
  - Documenta los mejores parámetros encontrados.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

- **Evaluación**:
  - Usa métricas relevantes para regresión (e.g., RMSE, MAE, R^2) y explica su significado en el contexto agrícola.
  - Incluye visualizaciones de resultados (e.g., gráfico de predicciones vs. valores reales, residuos).
- **Interpretación**: Explica cómo las variables (e.g., pH, temperatura) influyen en las predicciones (puedes usar SHAP o coeficientes del modelo).

3.3. Visualizaciones y resultados

- **Gráficos relevantes**:
  - Incluye visualizaciones del EDA (e.g., matriz de correlación, distribución de variables).
  - Muestra gráficos de evaluación del modelo (e.g., curvas de aprendizaje, dispersión de predicciones).
- **Resumen de resultados**: Redacta un apartado con las conclusiones clave, como el rendimiento del modelo y su utilidad práctica en agricultura de precisión.
- **Exportación de resultados**: Si es necesario, guarda los resultados (e.g., predicciones, métricas) en un archivo (CSV, JSON) o visualización exportada (PNG, PDF

3.4 Registro de Experimentos

Mantener una sección o tabla que documente los experimentos realizados:

Experimento	Configuración	Resultado (métricas principales)	observaciones
Exp1	batch = 32, lr = 0.001	Precisión: 87.3%; Loss = 650.8% ; F1=….	Convergencia lenta
Exp2	batch = 36, lr = 0.002	Precisión: 89.1%; Loss= 55.4%; F1=….	Mejor resultado pero inestable
:			
Expn			

3.5 Reproducibilidad y robustez

- Documentar cambios significativos en el notebook
- Incluir referencias a problemas resueltos
- Dependencias:
  - Incluir una celda al inicio del notebook con todas las importaciones necesarias (e.g., numpy, pandas, scikit-learn, matplotlib).
  - Especificar las versiones de las bibliotecas usadas (e.g., !pip install scikit-learn==1.5.1).
  - Control de versiones: Si se usa un dataset externo, documenta su fuente y versión (o guárdalo en un lugar accesible como Google Drive).
  - Pruebas de robustez:
    - Verificar que el notebook se ejecute de principio a fin sin errores (usa "Reiniciar y ejecutar todo" en Colab).
    - Probar con un subconjunto de datos o en un entorno diferente para asegurar portabilidad.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

4. Consideraciones Adicionales para la Auto documentación

4.1 Consideraciones éticas y prácticas

- Utilizar nombres descriptivos para variables y funciones
- Mantener consistencia en la nomenclatura
- Validación del dominio: Asegurarse que las predicciones sean realistas para el contexto agrícola (e.g., rendimientos negativos no tienen sentido).
- Limitaciones: Documentar las limitaciones del modelo (e.g., falta de datos de ciertas regiones, suposiciones sobre el clima).
- Impacto práctico: Explicar cómo el modelo podría usarse en la agricultura de precisión (e.g., optimización de riego, fertilización).

4.2 Referencias y Citas

- Incluir referencias bibliográficas para algoritmos, técnicas o implementaciones utilizadas
- Citar adecuadamente el código adaptado de otras fuentes

4.3 Conclusiones y Trabajo Futuro

Cada notebook debe finalizar con:

- Resumen de logros
- Limitaciones identificadas
- Próximos pasos sugeridos

5. Lista de Verificación General

- **Revisión de errores**: Verificar que no haya celdas con errores o salidas incompletas.
- **Ortografía y gramática**: Corregir cualquier error en los textos de Markdown.
- **Optimización**: Eliminar celdas innecesarias o código redundante.
- **Prueba de audiencia**: Si es posible, pedir a un integrante del grupo o externo, que revise el *notebook* para confirmar su claridad.

5. Notas adicionales

- **Contexto agrícola**: Asegúrarse que las métricas y resultados sean interpretables en términos de agricultura (e.g., "un RMSE de 0.5 toneladas/ha indica un error aceptable para la planificación").
- **Eficiencia en Colab**: Si el dataset es grande, considerar usar submuestras durante el desarrollo para ahorrar recursos.
- **Interactividad**: En Colab, aprovechar widgets interactivos (e.g., ipywidgets) para explorar parámetros o resultados.

6. Lista de Verificación Final

Antes de dar por finalizado cada notebook, verificar:

- [ ] ¿El notebook tiene un título claro y los nombres de los autores?
- [ ] ¿La estructura sigue el esquema recomendado?
- [ ] ¿Todas las celdas de código están documentadas adecuadamente?
- [ ] ¿Se han documentado las decisiones importantes y sus justificaciones?
- [ ] ¿Las visualizaciones tienen títulos, etiquetas y leyendas apropiadas?
- [ ] ¿Se incluyen interpretaciones para los resultados presentados?

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025

- [ ] ¿El código es reproducible (semillas fijadas, dependencias documentadas)?
- [ ] ¿Se han citado correctamente las fuentes externas?
- [ ] ¿Se incluyen conclusiones y recomendaciones para trabajo futuro?

Esta estructura de auto documentación completa elimina la necesidad de documentación externa impresa, ya que cada notebook contiene toda la información necesaria para entender el desarrollo, implementación y evaluación del sistema de “Predicción de Rendimiento de Cultivos en Agricultura de Precisión a Partir de Condiciones del Suelo y Clima”.

Material exclusivo de la Asignatura - Inteligencia Artificial. – Dr. Oscar Magna – UTEM 2025