PLAN DE TRABAJO

Monografía IB — Informática / Computer Science

Predicción de Diabetes Tipo 2: Deep Learning vs ML Clásico

Deep Learning (MLP) vs ML Clásico (Random Forest)

Estudiante: Natalia
Tutor: Especialista en ML/IA
Duración estimada: 22 semanas

Pregunta de investigación definitiva

> ¿En qué medida un modelo de deep learning (MLP) supera al ML clásico (Random Forest) en la predicción de diabetes tipo 2, evaluado mediante accuracy, F1-score y AUC-ROC, utilizando el dataset Pima Indians Diabetes?

Desglose de la pregunta

- **Qué se compara:** Deep Learning (MLP — red neuronal multicapa) vs ML Clásico (Random Forest — ensemble de árboles de decisión)
- **Tarea específica:** Predicción de diabetes tipo 2 (clasificación binaria: diabético / no diabético)
- **Métricas concretas:** Accuracy (precisión global), F1-score (balance precisión-recall), AUC-ROC (rendimiento global del clasificador)
- **Dataset específico:** Pima Indians Diabetes — 768 registros, 8 variables clínicas, disponible en Kaggle/UCI
- **Exige análisis:** La pregunta no se responde con sí/no. Requiere medir, comparar, e interpretar POR QUÉ.

> 🎯 **Por qué esta pregunta cumple con el IB**
>
> Criterio A: Es específica, medible, y determina la metodología (experimental comparativa).
>
> Criterio B: Exige explicar qué es deep learning (MLP), qué es ML clásico (RF), cómo funcionan, por qué se eligieron, y cuál es el debate actual entre ambos paradigmas.
>
> Criterio C: El análisis de múltiples métricas + contexto médico + debate DL vs ML clásico garantiza profundidad.
>
> Criterio D: La estructura fluye naturalmente: marco → método → resultados → discusión.
>
> Criterio E: El proceso de aprender ML desde cero es material riquísimo para reflexión.

Estructura de la monografía

Esta es la estructura sección por sección con el conteo de palabras objetivo y el contenido específico que debe ir en cada parte. Meta total: 3.800-3.950 palabras (el límite es 4.000).

Sección	Palabras	Contenido específico
Página de título	No cuenta	Título, pregunta de investigación, asignatura, conteo de palabras

Abstract / Resumen	Max 300	Resumen del trabajo: pregunta, método, conclusión. Se escribe AL FINAL.
Índice	No cuenta	Secciones con números de página
Introducción	300-400	Contexto (diabetes + ML), justificación, pregunta de investigación, alcance y limitaciones
Marco teórico	600-800	¿Qué es ML? ¿Qué es deep learning? ¿Qué es un MLP y por qué es deep learning? ¿Qué es Random Forest y por qué es ML clásico? Diferencias entre ambos paradigmas. ¿Qué dice la literatura sobre DL vs ML clásico en diabetes? Métricas: accuracy, F1, AUC-ROC
Metodología	400-500	Dataset (origen, variables, limitaciones), preprocesamiento (normalización, valores nulos, SMOTE), división train/test, hiperparámetros, herramientas (Python, versiones, librerías)
Resultados	600-800	Tablas comparativas de métricas, matrices de confusión, curvas ROC, gráficos de importancia de variables. Los gráficos NO cuentan para el límite.
Análisis y discusión	800-1000	Interpretación: ¿por qué un modelo superó al otro? ¿El deep learning justifica su complejidad frente al ML clásico con este dataset? Falsos negativos vs falsos positivos en contexto médico. Comparación con literatura. Limitaciones del estudio. Interpretabilidad (RF como ML clásico permite ver importancia de variables; MLP como deep learning es caja negra).
Conclusión	200-300	Síntesis (NO repetir resultados), respuesta a la pregunta, limitaciones, preguntas abiertas, futuras líneas de investigación
Bibliografía	No cuenta	Mínimo 10-15 fuentes académicas. Formato APA 7. Incluir papers, libros, documentación técnica.
Anexos	No cuenta	Código fuente completo (Python), tablas de datos brutos, gráficos adicionales, hiperparámetros probados

> ⚠ **Recordatorio crítico**
>
> El examinador NO lee nada después de la palabra 4.000. Las tablas de datos, gráficos, bibliografía y anexos NO cuentan. Apuntar a 3.800-3.950 para tener margen. El abstract se escribe AL FINAL, cuando todo el trabajo está terminado.

Cronograma semana a semana

Este cronograma de 22 semanas integra tres líneas paralelas: el aprendizaje de ML, la investigación/experimentación, y la redacción. Las estrellas (☆) marcan hitos críticos y reflexiones RPPF obligatorias.

Semana	Fase	Actividades	Entregable
1	ARRANQUE	Confirmar asignatura CS con coordinador IB. Instalar Python, Jupyter, librerías (Pandas, Scikit-learn, Keras, Matplotlib, Seaborn). Descargar dataset Pima Indians Diabetes de Kaggle.	Entorno listo, dataset descargado
2	INVESTIGACIÓN PRELIMINAR	Leer 3-4 artículos introductorios sobre ML aplicado a diabetes. Entender qué es el dataset Pima Indians: origen, variables, contexto médico. Instalar Zotero y empezar a guardar fuentes.	Carpeta de fuentes, notas de lectura
3	APRENDIZAJE ML #1	Curso básico: ¿Qué es Machine Learning? Tipos (supervisado, no supervisado). ¿Qué es clasificación? Primeros pasos con Pandas: cargar CSV, explorar datos (.head(), .describe(), .info()).	Primer notebook: cargar y explorar el dataset
4	APRENDIZAJE ML #2	EDA completo: distribuciones, correlaciones, valores faltantes, boxplots. Visualización con Matplotlib/Seaborn. Entender desbalance de clases (más sanos que diabéticos en el dataset).	Notebook EDA con gráficos
5	FORMULACIÓN PREGUNTA	Redactar pregunta de investigación definitiva. Crear outline de la monografía. Definir cronograma personal. Preparar primera reflexión RPPF.	Pregunta + outline + cronograma
5-6	☆ REFLEXIÓN RPPF #1	Reunión con supervisor: presentar tema, pregunta, plan. Documentar en RPPF: ¿por qué este tema? ¿Qué espero encontrar? ¿Qué desafíos anticipo? ¿Qué he aprendido hasta ahora?	RPPF #1 completado
6	APRENDIZAJE ML #3	Preprocesamiento: normalización (MinMaxScaler, StandardScaler), manejo de valores nulos/cero, división train/test (80/20), concepto de validación cruzada.	Notebook preprocesamiento
7	APRENDIZAJE ML #4	Primer modelo: KNN (como calentamiento). Entender distancia	Notebook KNN funcionando

		euclidiana, efecto de K, métricas básicas (accuracy, matriz de confusión). Implementar con Scikit-learn.	
8	**APRENDIZAJE ML #5**	Segundo modelo: Random Forest. Entender árboles de decisión, bagging, importancia de variables (feature importance). Implementar y comparar con KNN.	Notebook Random Forest
9	**APRENDIZAJE ML #6**	Tercer modelo: Red Neuronal MLP con Keras. Entender neuronas, capas, funciones de activación (ReLU, sigmoid), épocas, batch size. Implementar MLP básico.	Notebook MLP básico
10	**APRENDIZAJE ML #7**	Métricas avanzadas: Precisión, Recall, F1-score, AUC-ROC. ¿Por qué accuracy sola no basta? Falsos negativos en medicina. Implementar todas las métricas para los 3 modelos.	Notebook comparativo con todas las métricas
10	**☆ HITO: EXPERIMENTO BÁSICO COMPLETO**	En este punto Natalia ya tiene los 3 modelos entrenados y comparados con métricas básicas. Revisar resultados y decidir si el alcance es suficiente o si agrega optimización.	Resultados preliminares completos
11	**OPTIMIZACIÓN**	Optimizar hiperparámetros: GridSearchCV para Random Forest (n_estimators, max_depth). Ajustar arquitectura MLP (número de capas, neuronas, learning rate, epochs). Documentar todas las variaciones probadas.	Modelos optimizados
12	**ANÁLISIS PROFUNDO**	Validación cruzada 5-fold para robustez. Análisis de importancia de variables (Random Forest). Curvas ROC comparativas. Matrices de confusión finales con interpretación médica.	Todos los gráficos y tablas finales
13	**INVESTIGACIÓN BIBLIOGRÁFICA**	Leer 8-10 papers sobre ML aplicado a predicción de diabetes. Buscar en PubMed, Google Scholar, IEEE. Crear tabla comparativa: ¿qué algoritmos usaron? ¿Qué métricas obtuvieron? ¿Con qué dataset?	Tabla comparativa de literatura, 10+ fuentes
14	**☆ REFLEXIÓN RPPF #2**	Reunión intermedia con supervisor: revisar resultados, discutir hallazgos, ajustar rumbo. Documentar: ¿qué me sorprendió? ¿Qué cambié? ¿Qué dificultades encontré? ¿Cómo las resolví?	RPPF #2 completado
15-16	**REDACCIÓN: INTRO + MARCO**	Escribir introducción (300-400 palabras) y marco teórico (600-800 palabras). Explicar el debate deep learning vs ML	Borrador intro + marco teórico

		clásico, qué es MLP (deep learning), qué es Random Forest (ML clásico), con terminología precisa. Citar fuentes.	
17	**REDACCIÓN: METODOLOGÍA**	Escribir metodología (400-500 palabras). Documentar TODO: versiones de Python/librerías, pasos de preprocesamiento, hiperparámetros finales, random seed. Debe ser reproducible.	Borrador metodología
18	**REDACCIÓN: RESULTADOS**	Escribir resultados (600-800 palabras). Insertar tablas y gráficos (estos no cuentan para el límite). Presentar métricas de forma clara y organizada.	Borrador resultados
19	**REDACCIÓN: ANÁLISIS**	Escribir análisis y discusión (800-1000 palabras). Esta es la sección MAS IMPORTANTE. Interpretar POR QUÉ un paradigma superó al otro, discutir si deep learning justifica su complejidad frente a ML clásico, comparar con literatura, limitaciones, falsos negativos en contexto médico.	Borrador análisis
20	**REDACCIÓN: CONCLUSIÓN**	Escribir conclusión (200-300 palabras). Síntesis razonada, NO repetir resultados. Limitaciones, preguntas abiertas, futuras líneas. Escribir abstract (300 palabras max). Revisar bibliografía APA.	Primer borrador COMPLETO
21	**REVISIÓN Y PULIDO**	Revisión completa: ¿cada sección responde a la pregunta? ¿Hay análisis o solo descripción? Conteo de palabras (3.800-3.950 meta). Formato, citas, rotulado de tablas/gráficos. Preparar anexos con código.	Versión final de la monografía
22	**☆ REFLEXIÓN RPPF #3 + VIVA VOCE**	Reflexión final: ¿qué aprendí? ¿Qué haría diferente? ¿Cómo crecí como investigadora? Entrevista con supervisor (viva voce). Entrega final.	RPPF #3 + entrega final

Ruta de aprendizaje de Machine Learning

Esta ruta está diseñada para alguien con Python intermedio pero sin experiencia en ML. Cada semana tiene un concepto teórico y un ejercicio práctico. El tutor (especialista en ML/IA) guía la comprensión conceptual; Natalia implementa y experimenta.

Semana	Concepto ML	Lo que Natalia aprende	Ejercicio práctico
3	**¿Qué es Machine Learning?**	Diferencia entre programación tradicional y ML. Tipos: supervisado (clasificación, regresión) vs no supervisado. Por qué la predicción de diabetes es clasificación binaria.	Cargar el dataset con Pandas. Usar .head(), .describe(), .shape, .dtypes. Identificar la variable objetivo (Outcome).
4	**Análisis Exploratorio de Datos (EDA)**	Distribuciones (histogramas), correlaciones (heatmap), outliers (boxplots). Cómo los datos "cuentan una historia". Desbalance de clases y por qué importa.	Crear: histograma de glucosa por clase, heatmap de correlaciones, boxplot de IMC. Identificar valores sospechosos (glucosa=0, presión=0).
6	**Preprocesamiento de datos**	Por qué normalizar (diferentes escalas). MinMaxScaler vs StandardScaler. Manejo de valores faltantes/cero. Train/test split (80/20). Random state para reproducibilidad.	Reemplazar ceros sospechosos por medianas. Normalizar. Dividir datos. Guardar random_state=42.
7	**KNN (K-Nearest Neighbors)**	Concepto de distancia euclidiana. Cómo K define la "vecindad". Efecto de K: underfitting (K grande) vs overfitting (K pequeño). Accuracy y matriz de confusión.	Implementar KNN con Scikit-learn para K=3,5,7,9. Graficar accuracy vs K. Generar primera matriz de confusión.
8	**Random Forest**	Árboles de decisión: cómo "preguntan" sobre los datos. Bagging: muchos árboles votan. Feature importance: qué variables pesan más. Ventaja: interpretabilidad.	Implementar Random Forest. Graficar feature importance. Comparar accuracy con KNN.
9	**Red Neuronal MLP (Keras)**	Neurona artificial: pesos, bias, función de activación. Capas: input, hidden, output. ReLU vs Sigmoid. Epochs y batch	Construir MLP: Input(8) → Dense(16, relu) → Dense(8, relu) → Dense(1, sigmoid). Entrenar 100 epochs. Comparar con RF.

		size. Forward pass y backpropagation (conceptual).	
10	**Métricas de evaluación**	Accuracy: ¿cuántos acertó? Precisión: de los que dijo "diabetes", ¿cuántos realmente tenían? Recall: de los que tenían diabetes, ¿cuántos detectó? F1: balance entre precisión y recall. AUC-ROC: rendimiento global.	Calcular TODAS las métricas para los 3 modelos. Crear tabla comparativa. Generar curvas ROC superpuestas.
11	**Optimización de hiperparámetros**	GridSearchCV: probar combinaciones sistemáticamente. Validación cruzada: por qué un solo train/test split no es confiable. Cómo evitar overfitting.	GridSearch en RF: n_estimators=[50,100,200], max_depth=[5,10,None]. Ajustar MLP: probar diferentes arquitecturas y learning rates.
12	**Análisis avanzado e interpretación**	Validación cruzada 5-fold. Matrices de confusión en contexto médico: falso negativo = paciente con diabetes no detectado (peligroso). Curvas ROC comparativas.	Ejecutar 5-fold CV. Analizar: ¿qué modelo tiene mejor recall? ¿Cuál tiene menos falsos negativos? Generar todos los gráficos finales.

Recursos de aprendizaje recomendados

Curso gratuito: "Machine Learning Crash Course" de Google (en inglés, gratuito, práctico)

Curso gratuito: "Intro to Machine Learning" de Kaggle Learn (muy práctico, con notebooks)

Libro: "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" de Aurélien Géron (capítulos 1-6 y 10)

Documentación: Scikit-learn User Guide (https://scikit-learn.org/stable/user_guide.html)

Dataset: Pima Indians Diabetes en Kaggle (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)

Videos: StatQuest con Josh Starmer en YouTube (excelentes explicaciones visuales de ML)

Herramientas y configuración técnica

Software (todo gratuito)

- **Python 3:** Lenguaje principal. Versión recomendada: 3.10 o superior.
- **Jupyter Notebook:** Para crear y documentar notebooks paso a paso. Alternativa: Google Colab (nube, gratuito, no requiere instalación).
- **Pandas:** Manipulación de datos (DataFrames). Instalar: pip install pandas
- **NumPy:** Operaciones numéricas. Instalar: pip install numpy
- **Scikit-learn:** Random Forest, KNN, métricas, preprocesamiento, GridSearch. Instalar: pip install scikit-learn
- **TensorFlow / Keras:** Red neuronal MLP. Instalar: pip install tensorflow (incluye Keras)
- **Matplotlib + Seaborn:** Gráficos y visualización. Instalar: pip install matplotlib seaborn
- **Zotero:** Gestor bibliográfico gratuito. Descargar de zotero.org

> **💻 Instalación rápida (un solo comando)**
>
> pip install pandas numpy scikit-learn tensorflow matplotlib seaborn jupyter
>
> Alternativa sin instalar nada: usar Google Colab (colab.research.google.com). Funciona en el navegador, tiene todas las librerías preinstaladas, y es gratuito. Ideal si la laptop de Natalia tiene limitaciones de rendimiento.

Configuración del experimento (para reproducibilidad)

El IB exige que la metodología sea lo suficientemente detallada para que otro investigador pueda reproducir el experimento. Natalia debe documentar y reportar estos parámetros en la sección de metodología:

- **Semilla aleatoria:** random_state=42 en TODOS los modelos y en train_test_split. Esto garantiza que cualquiera que corra el código obtenga exactamente los mismos resultados.
- **División train/test:** 80% entrenamiento / 20% prueba. También usar validación cruzada 5-fold para robustez.
- **Normalización:** StandardScaler (media=0, desviación estándar=1). Justificación: las variables tienen escalas muy diferentes (glucosa: 0-200, edad: 21-81, embarazos: 0-17).
- **Valores faltantes:** Los ceros en glucosa, presión arterial, IMC, grosor de piel e insulina son biológicamente imposibles — son valores faltantes disfrazados. Reemplazar por la mediana de cada variable.

- **Hiperparámetros a reportar:** n_estimators, max_depth, min_samples_split para RF. Número de capas, neuronas, learning rate, epochs, batch_size para MLP. Documentar los valores finales Y los probados.
- **Métricas a reportar:** accuracy, precision, recall, F1-score, AUC-ROC, y matrices de confusión.

Guía para las reflexiones RPPF

El RPPF (Researcher's Planning and Progress Form) es donde Natalia documenta su proceso en tres momentos. El Criterio E (Compromiso) se evalúa exclusivamente a través de este formulario. Los examinadores buscan voz propia y crecimiento genuino.

☆ Reflexión #1 (Semana 5-6) — Inicio

Preguntas guía para Natalia:
• ¿Por qué elegí este tema? ¿Qué me motiva de la IA aplicada a salud?
• ¿Qué sé hasta ahora sobre ML? ¿Qué me falta aprender?
• ¿Qué espero encontrar en mi investigación? ¿Tengo hipótesis?
• ¿Qué desafíos anticipo? ¿Cómo pienso abordarlos?
• ¿Cómo se conecta este proyecto con mis metas futuras (CS, universidad, beca)?

☆ Reflexión #2 (Semana 14) — Intermedia

Preguntas guía para Natalia:
• ¿Qué me ha sorprendido de los resultados hasta ahora?
• ¿Tuve que cambiar algo de mi plan original? ¿Por qué?
• ¿Qué dificultades técnicas encontré y cómo las resolví?
• ¿Mi pregunta de investigación sigue siendo apropiada o necesita ajuste?
• ¿Qué estoy aprendiendo sobre el proceso de investigación en sí mismo?

☆ Reflexión #3 (Semana 22) — Final

Preguntas guía para Natalia:
• ¿Qué aprendí sobre Machine Learning que no sabía al comenzar?
• ¿Qué aprendí sobre el proceso de investigación científica?
• Si pudiera empezar de nuevo, ¿qué haría diferente?
• ¿Cómo cambió mi comprensión de la IA en salud a lo largo de este proyecto?
• ¿Cómo me preparó esto para mi futuro en Computer Science?

Estrategia de beca: cómo aprovechar esta monografía

La monografía no termina cuando se entrega al IB. Si se ejecuta bien, se convierte en una pieza central de las aplicaciones de beca y admisión universitaria. Aquí está cómo aprovecharla:

- **Ensayos de admisión:** En los ensayos de admisión (Common App, UCAS, etc.), Natalia puede describir cómo su monografía la llevó a descubrir su pasión por la IA en salud. La narrativa es: "Comparé deep learning vs ML clásico para predecir diabetes en mi región, y descubrí que la IA más compleja no siempre es la mejor solución."
- **Portafolio técnico:** Publicar el código en GitHub con documentación clara. Esto demuestra habilidad técnica Y capacidad de comunicar. Incluir el link en aplicaciones.
- **Carta de recomendación:** Si la monografía obtiene A o B, mencionarlo explícitamente. Muchas universidades reconocen el rigor de la monografía IB.
- **Ángulo de impacto social:** El ángulo "detección temprana de diabetes en contextos con acceso limitado a especialistas en Latinoamérica" es atractivo para becas que valoran impacto social (Fulbright, MasterCard Foundation, Open Society, becas institucionales de universidades).
- **Presentación en ferias:** Si los resultados son interesantes, considerar presentar en ferias de ciencia locales o nacionales (Expociencia, por ejemplo). Eso agrega una línea más al CV académico.

— Fin del plan de trabajo —

¡Manos a la obra, Natalia! Semana 1 comienza ahora.