UNIVERSIDAD TECNOLÓGICA METROPOLITANA
UTEM del Estado de Chile

INFB6052 - HERRAMIENTAS PARA CS. DE DATOS
INGENIERÍA CIVIL EN CIENCIA DE DATOS

Informe de Anteproyecto:

Propuesta de un pipeline de herramientas de Machine Learning para el diagnóstico de cáncer de mama, comparando MLP y SVM.

Ignacio Ramírez
Cristian Vergara
Antonia Montecinos
Grupo 2

Segundo Semestre de 2025

Propuesta de Anteproyecto:

(i) Estructura del Anteproyecto

El presente documento sigue la estructura requerida para la propuesta de anteproyecto, la cual se organiza en las siguientes secciones:

- **Título Tentativo:** Nombre que encapsula el objetivo principal del proyecto.
- **Justificación:** Relevancia del problema y su alineación con el enfoque en herramientas de Machine Learning.
- **Objetivos:** Metas generales y específicas orientadas a la construcción de un pipeline completo.
- **Metodología Propuesta:** Descripción de los pasos, desde la gestión de datos hasta el despliegue.
- **Estado del Arte Breve:** Análisis de trabajos previos para contextualizar el problema.
- **Cronograma Tentativo:** Planificación temporal de las actividades del proyecto.
- **Referencias:** Listado de fuentes académicas y técnicas citadas.

(ii) Título Tentativo del Proyecto

Pipeline de Herramientas de Machine Learning para el Diagnóstico de Cáncer de Mama: Desde la Gestión de Datos hasta el Despliegue de un Prototipo Interactivo.

(iii) Justificación

El diagnóstico temprano y preciso del cáncer de mama es crucial para mejorar los resultados de los tratamientos. El Machine Learning ofrece herramientas poderosas para asistir a los profesionales médicos en esta tarea, analizando datos clínicos para predecir si un tumor es benigno o maligno. Este proyecto se centra en la construcción de un pipeline de herramientas de ciencia de datos robusto y completo para abordar este problema, en lugar de enfocarse únicamente en la complejidad del modelo predictivo.
Se utilizará el dataset "Breast Cancer Wisconsin (Diagnostic)", que contiene características numéricas extraídas de imágenes de biopsias. Este enfoque permite trabajar con datos tabulares estructurados, ideales para comparar el rendimiento de diferentes familias de modelos de clasificación, como las redes neuronales (MLP) y los modelos basados en márgenes (SVM).
La propuesta está directamente alineada con los objetivos del curso "Herramientas para Cs. de Datos". El énfasis se pone en el dominio práctico del ecosistema de herramientas: la gestión de datos con MongoDB, el preprocesamiento y modelado con Scikit-learn, y el despliegue de un prototipo interactivo con Streamlit. El resultado será un sistema funcional de principio a fin, que demuestra un flujo de trabajo de MLOps elemental y reproducible para un problema médico relevante.

(iv) Objetivos Generales y Específicos

Objetivo General

Desarrollar un pipeline completo y un prototipo funcional para el diagnóstico de cáncer de mama, abarcando desde la gestión de datos tabulares y el entrenamiento comparativo de modelos de clasificación, hasta el despliegue de una interfaz interactiva de predicción.

Objetivos Específicos

1. **Implementar un sistema de gestión de datos** utilizando MongoDB para almacenar los registros de pacientes y sus características clínicas del dataset, facilitando un acceso estructurado.

2. **Desarrollar un pipeline de preprocesamiento de datos**, que incluya el escalado de características (ej. StandardScaler) para preparar los datos numéricos para el entrenamiento de los modelos.

3. **Entrenar y comparar el rendimiento de dos modelos de clasificación:** un Perceptrón Multicapa (MLP) y una Máquina de Vectores de Soporte (SVM), para seleccionar el modelo más performante para el despliegiegue.

4. **Desarrollar y desplegar una aplicación web interactiva con Streamlit** que permita al usuario ingresar los valores de las características de un nuevo caso y obtener una predicción diagnóstica en tiempo real del mejor modelo entrenado.

(v) Metodología Inicial Propuesta

El proyecto se estructurará en módulos secuenciales, cada uno enfocado en una etapa del pipeline de Machine Learning.

- **Fuente de Datos:** Se utilizará el dataset público **Breast Cancer Wisconsin (Diagnostic)** del repositorio de UCI Machine Learning.
  - **URL:** https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
  - **Descripción:** Contiene 569 registros y 30 características numéricas que describen las propiedades de los núcleos celulares. El objetivo es una clasificación binaria (maligno/benigno).
- **1. Gestión de Datos:**
  - **Herramienta:** MongoDB (con ‘pymongo‘).
  - **Proceso:** Cargar el dataset desde un archivo CSV a una colección en MongoDB. Cada documento representará un caso, almacenando sus 30 características y la etiqueta de diagnóstico.
- **2. Preparación y Análisis de Datos:**
  - **Herramientas:** Pandas, Scikit-learn.
  - **Proceso:** Se leerán los datos desde MongoDB a un DataFrame de Pandas. Se dividirán los datos en conjuntos de entrenamiento y prueba. Se aplicará un escalado estándar (‘StandardScaler‘) a las características para normalizar sus rangos, un paso crucial para el buen rendimiento tanto del MLP como del SVM.
- **3. Modelado y Evaluación:**
  - **Herramienta:** Scikit-learn.
  - **Proceso:** Se implementará un ‘MLPClassifier‘ y un ‘SVC‘ (Support Vector Classifier). Ambos modelos se entrenarán con los mismos datos de entrenamiento escalados. Se utilizará validación cruzada para una comparación robusta de su rendimiento. La evaluación final se realizará sobre el conjunto de prueba utilizando métricas como exactitud, precisión, recall, F1-score y la matriz de confusión. El modelo con el mejor desempeño global será seleccionado.
- **4. Despliegue Interactivo:**
  - **Herramienta:** Streamlit.
  - **Proceso:** Se creará una aplicación web con una interfaz que presente 30 campos de entrada (sliders o inputs numéricos). El usuario podrá ingresar valores, y la aplicación aplicará el mismo escalado y usará el modelo final seleccionado (MLP o SVM) para generar una predicción ("Maligno" o "Benigno") y mostrarla en pantalla.

(vi) Estado del Arte Breve

El dataset "Breast Cancer Wisconsin (Diagnostic)" es un benchmark estándar y ampliamente utilizado en la comunidad de Machine Learning para probar y comparar algoritmos de clasificación. La literatura académica presenta numerosos estudios que lo utilizan para validar la eficacia de diferentes modelos. Un ejemplo representativo es el trabajo de Chaurasia y Pal (2017) [1], quienes comparan el rendimiento de SMO (un algoritmo para SVM), IBK (k-NN) y árboles de decisión. Su estudio concluyó que el algoritmo basado en SVM obtenía la mayor precisión para la clasificación de los datos. De manera similar, Athar e Ilavarasi (2020) [2] realizaron un estudio comparativo que incluía Regresión Logística y SVM, incorporando además métodos de selección de características para optimizar el rendimiento.
Análisis Crítico y Diferenciación: Estos estudios, y muchos otros similares, son fundamentales para establecer líneas base de rendimiento, pero su objetivo principal es la **optimización de métricas para encontrar el "mejor" modelo clasificador**. Nuestra propuesta, en cambio, aborda el problema desde la perspectiva de la **ingeniería de software y las herramientas de datos**. No buscamos superar el estado del arte en precisión, sino construir un **sistema completo, reproducible y desplegable**. El valor de nuestro trabajo radica en la integración de MongoDB para la gestión de datos, la creación de un pipeline de preprocesamiento robusto y, crucialmente, el despliegue de un prototipo interactivo con Streamlit que materializa el ciclo de vida completo de un proyecto de Machine Learning. Este enfoque en la aplicación práctica de MLOps elementales es el principal diferenciador de nuestra propuesta.

(vii) Cronograma Tentativo

Se propone un cronograma de 8 semanas para el desarrollo del proyecto, abarcando desde el 9 de octubre hasta la fecha de entrega final el 4 de diciembre.

Semanas	**Actividad / Hito Entregable**
Semanas 1-2 (9 Oct - 22 Oct)	**Fase de Datos y Entorno:** - Configuración del entorno (MongoDB, Python). - Creación del script para cargar el dataset de cáncer de mama en MongoDB. **Hito 1:** Base de datos poblada y funcional.
Semanas 3-4 (23 Oct - 5 Nov)	**Preprocesamiento y Modelado Comparativo:** - Desarrollo del script de preprocesamiento (lectura, escalado, división). - Implementación y entrenamiento inicial del MLP y del SVM. **Hito 2:** Primeros modelos entrenados con resultados preliminares.
Semanas 5-6 (6 Nov - 19 Nov)	**Evaluación y Selección de Modelo:** - Evaluación exhaustiva y comparación de ambos modelos con métricas de clasificación. - Ajuste de hiperparámetros y selección del modelo final. - Guardado del modelo seleccionado (ej. usando ‘joblib‘).
Semana 7 (20 Nov - 26 Nov)	**Despliegue del Prototipo:** - Desarrollo de la interfaz de usuario en Streamlit con los campos de entrada. - Integración del modelo final para realizar predicciones en tiempo real. **Hito 3:** Prototipo funcional para pruebas.
Semana 8 (27 Nov - 3 Dic)	**Fase Final y Documentación:** - Limpieza del código, organización del repositorio y redacción del informe final. - Preparación de la presentación. **Entrega Final (4 de Diciembre).**

Cuadro 1: Cronograma de actividades ajustado (8 semanas).

Referencias

[1] Chaurasia, V., & Pal, S. (2017). *A Novel Approach for Breast Cancer Detection using Data Mining Techniques*. International Journal of Innovative Research in Computer and Communication Engineering, 5(1), 136-143. Disponible en: https://ssrn.com/abstract=2994932

[2] Athar, A., & Ilavarasi, A. K. (2020). *A Comparative study of machine learning models for breast cancer prediction*. Journal of Physics: Conference Series, 1716, 012052. DOI: 10.1088/1742-6596/1716/1/012052