MDI: Plataforma de Diseño Inteligente de Materiales Cementíceos

Presentación de avance de trabajo

---

1. Contexto y motivación

En el marco de mi transición de la maestría en ciencia de datos hacia el doctorado en materiales, desarrollé una plataforma funcional que conecta ambas disciplinas: **MDI (Material Design Intelligence)**, un sistema web que aplica técnicas de machine learning e inteligencia artificial explicable al problema del diseño de mezclas de concreto.

La motivación surge de una observación directa del campo: el diseño de mezclas cementíceas sigue siendo, en gran parte, un proceso basado en prueba y error experimental, tablas empíricas y experiencia acumulada. Cada iteración en el laboratorio tiene un costo de tiempo y recursos. La relación entre los componentes de una mezcla (cemento, ceniza volante, agua, aditivos, agregados) y la resistencia a compresión resultante es altamente no lineal, con interacciones complejas entre variables que dificultan la optimización manual.

La pregunta que guía este trabajo es: **¿se puede construir un asistente computacional que, a partir de datos existentes, permita al ingeniero de materiales explorar el espacio de diseño de forma informada, entendiendo no solo qué predice el modelo sino por qué lo predice?**

---

2. Qué es MDI y qué hace

MDI es una aplicación web compuesta por un backend en Python (FastAPI, scikit-learn, SHAP) y un frontend en React/TypeScript. No es un sistema teórico ni un notebook aislado: es una plataforma interactiva donde el usuario puede realizar el ciclo completo de trabajo sin escribir código.

El flujo de uso es el siguiente:

1. **Explorar los datos** — ver las distribuciones de cada variable, las correlaciones entre componentes, y entender la estructura del dataset antes de modelar.
2. **Entrenar un modelo predictivo** — el sistema entrena un Random Forest o Gradient Boosting sobre datos reales de resistencia a compresión de concreto. El modelo aprende la función:

   f(cemento, ceniza volante, agua, superplastificante, agregados, edad) □ *resistencia (MPa)*
3. **Entender por qué predice lo que predice** — usando SHAP (SHapley Additive exPlanations), el sistema descompone cada predicción en las contribuciones individuales de cada componente.
4. **Explorar el espacio de diseño** — variar ingredientes de forma sistemática para encontrar configuraciones óptimas.
5. **Guardar y exportar las mezclas candidatas** — para llevarlas al laboratorio.

Lo importante es que estos cinco pasos están integrados en una única plataforma, no son scripts separados.

3. Los datos

El sistema trabaja con tres datasets:

- **UCI Concrete Compressive Strength** (1,030 muestras, dataset real del repositorio UCI Machine Learning). Contiene 8 variables de entrada: cemento, escoria de alto horno, ceniza volante, agua, superplastificante, agregado grueso, agregado fino y edad de curado. La variable de salida es la resistencia a compresión en MPa.
- **ConcreteXAI** (500 muestras, sintético). Dataset extendido con 7 variables, generado con relaciones no lineales realistas para ampliar el análisis de explicabilidad.
- **Geopolymer** (400 muestras, sintético). Simula concreto geopolimérico con alto contenido de ceniza volante, donde la química de la mezcla es diferente al concreto Portland convencional.

Los datasets sintéticos se generaron con funciones que capturan relaciones conocidas en la literatura: la relación agua/cementante como factor dominante, la contribución logarítmica de la edad de curado, y el efecto positivo del superplastificante en la trabajabilidad y resistencia.

Las tres fuentes pueden unificarse en un conjunto combinado de 1,930 muestras con 7 variables comunes, lo que permite análisis comparativos entre tipos de concreto.

4. Los modelos: cómo funcionan realmente

4.1 Random Forest

El Random Forest es un ensemble de árboles de decisión. Cada árbol se entrena sobre un subconjunto aleatorio de los datos y un subconjunto aleatorio de las variables. Para hacer una predicción, los 100 árboles (valor por defecto) generan cada uno su propia estimación de la resistencia, y el resultado final es el promedio.

¿Por qué funciona bien para este problema? Porque las relaciones entre componentes y resistencia son no lineales y con interacciones. Un árbol individual puede capturar que “si el cemento es alto Y el agua es baja, la resistencia es alta”, pero un solo árbol sobreajusta. Al promediar 100 árboles entrenados con diferentes subconjuntos, se reduce la varianza sin perder la capacidad de capturar no linealidades.

En los datos de concreto, el Random Forest alcanza un R^2 superior a 0.90, lo que significa que explica más del 90% de la variabilidad en la resistencia a compresión.

4.2 Gradient Boosting

El Gradient Boosting construye árboles de forma secuencial: cada nuevo árbol se entrena sobre los errores residuales del modelo acumulado anterior. Es decir, el primer árbol hace una predicción aproximada, el segundo corrige lo que el primero se equivocó, el tercero corrige lo que los dos anteriores no pudieron, y así sucesivamente.

Tiene un parámetro adicional importante: la **tasa de aprendizaje** (learning rate), que controla cuánto "confía" el modelo en cada nuevo árbol. Una tasa baja requiere más árboles pero generaliza mejor.

4.3 Qué datos da el modelo

Para cualquier mezcla que el usuario ingrese, el sistema devuelve:

- **Predicción puntual**: por ejemplo, 34.5 MPa.
- **Intervalo de confianza al 95%**: por ejemplo, [29.1 — 39.9 MPa]. Esto le dice al ingeniero que el modelo no está seguro de un número exacto, sino que la resistencia real probablemente caerá en ese rango.
- **Desviación estándar**: una medida de cuánto "discrepan" las estimaciones internas del modelo.

El intervalo de confianza es fundamental para la toma de decisiones en ingeniería. No es lo mismo que el modelo diga "35 MPa [33-37]" (alta confianza, intervalo estrecho) que "35 MPa [20-50]" (baja confianza, mucha incertidumbre). El segundo caso indica que el modelo no tiene suficiente información para esa región del espacio de diseño, y el ingeniero debería ser más cauteloso.

4.4 Cómo se calcula la incertidumbre

Para Random Forest, la incertidumbre se calcula directamente a partir de la dispersión entre los árboles individuales. Si tengo 100 árboles y 95 de ellos predicen entre 30 y 40 MPa, pero 5 predicen valores muy diferentes, el intervalo del percentil 2.5 al 97.5 captura esa dispersión.

Para Gradient Boosting, se usan **modelos de regresión cuantil**: además del modelo principal (que predice la media), se entrenan dos modelos adicionales que predicen directamente el cuantil 5% y el cuantil 95% de la distribución condicional de la resistencia. Esto da un intervalo de confianza directo sin asumir distribución gaussiana.

---

5. Explicabilidad: por qué predice lo que predice

Uno de los aportes principales de este trabajo es la integración de SHAP (SHapley Additive exPlanations) como capa de explicabilidad.

5.1 Qué es SHAP

SHAP proviene de la teoría de juegos cooperativos. La idea es: dado que el modelo predijo 35 MPa para una mezcla particular, ¿cuánto contribuyó cada ingrediente a esa predicción?

El sistema descompone cada predicción de la siguiente forma:

```
Predicción = valor base + contribución(cemento) + contribución(agua) + contribución(ceniza) + …
```

El **valor base** es la predicción promedio del modelo (lo que predecería sin saber nada de la mezcla, aproximadamente 35.8 MPa para este dataset). Cada contribución puede ser positiva (aumenta la resistencia) o negativa (la disminuye).

5.2 Ejemplo concreto

Para una mezcla con 400 kg/m³ de cemento, 220 kg/m³ de agua y 28 días de curado:

```
Valor base:                  35.8 MPa
+ cement = 400 kg/m³:        +6.2 MPa   (más cemento → más resistencia)
+ water = 220 kg/m³:         −4.1 MPa   (más agua → menos resistencia)
+ age = 28 días:             +1.8 MPa   (más edad → más resistencia)
+ superplasticizer = 8:      +0.3 MPa
+ fly_ash = 50:              +0.2 MPa
+ coarse_aggregate = 1000:   −0.1 MPa
+ fine_aggregate = 700:      −0.1 MPa
+ blast_furnace_slag = 0:    +0.0 MPa

Predicción final:            40.0 MPa
```

Esto es enormemente valioso porque el ingeniero no solo sabe **qué** pasará, sino **por qué**. Si quiere subir la resistencia, puede ver que reducir el agua tiene un efecto negativo grande (-4.1 MPa con 220 kg/m³), lo que sugiere que reducir el agua sería una vía efectiva.

5.3 Tipos de análisis SHAP disponibles

- **Feature Importance**: ranking de qué variables son más influyentes en general (promedio absoluto de SHAP sobre todas las muestras).
- **Summary Plot**: visualización de todas las muestras donde se ve, para cada variable, cómo valores altos o bajos afectan positiva o negativamente la predicción.
- **Waterfall Plot**: la descomposición paso a paso de una predicción individual (como el ejemplo de arriba).
- **Dependence Plot**: cómo varía la contribución de una variable a lo largo de su rango, con interacciones detectadas automáticamente.

---

6. Exploración del espacio de diseño

6.1 Barrido paramétrico (1D)

El ingeniero define una “mezcla base” (por ejemplo, una receta estándar) y selecciona una variable a variar. El sistema genera predicciones + intervalos de confianza a lo largo de todo el rango.

Ejemplo: fijar todo constante y variar ceniza volante de 0 a 200 kg/m³. El resultado es una curva que muestra: - Cómo cambia la resistencia predicha - La banda de incertidumbre (zona sombreada al 95%) - La **región óptima** resaltada en verde (valores que dan ≥90% de la mejor predicción)

Esto responde directamente a preguntas como: “¿cuánta ceniza volante puedo agregar antes de que la resistencia caiga significativamente?”

6.2 Superficie de respuesta (2D)

Permite variar dos ingredientes simultáneamente y ver un mapa de calor de predicciones. Por ejemplo, variar ceniza volante (eje X) y agua (eje Y) da una superficie donde las zonas rojas son

alta resistencia y las amarillas son baja resistencia.

Esto permite identificar visualmente las combinaciones óptimas y entender cómo interactúan dos variables.

6.3 Comparación de configuraciones

Permite poner dos o más mezclas lado a lado y obtener para cada una: - Predicción con intervalo de confianza - Los 3 factores más influyentes según SHAP - Un gráfico comparativo

Esto es útil para la toma de decisiones: "la mezcla A da 38 MPa [34-42] y la mezcla B da 35 MPa [33-37]. La B tiene menor resistencia pero más certeza."

6.4 Visualización de microestructura

Como herramienta complementaria, el sistema genera una **sección transversal 2D procedural** del concreto basada en la composición. No es una imagen real sino una representación visual que muestra:

- Agregados gruesos (polígonos irregulares con zona de transición interfacial)
- Arena (partículas pequeñas)
- Ceniza volante (esferas perfectas oscuras, su forma característica vítrea)
- Poros (cuya cantidad depende de la relación agua/cementante)
- Pasta de cemento (cuyo color varía con la edad y contenido)

Incluye presets de mezclas típicas y un panel con propiedades derivadas (relación w/c, fracción de agregados, porcentaje de reemplazo). Sirve como herramienta de comunicación para visualizar lo que los números representan en la estructura real del material.

---

7. Gestión de configuraciones

Las mezclas que resultan prometedoras durante la exploración se pueden guardar con un nombre, su predicción y su intervalo de confianza. Después se pueden:

- Marcar como "candidatas a validación experimental"
- Exportar a CSV para llevar al laboratorio

El CSV incluye todas las columnas necesarias: cada ingrediente, la predicción del modelo, los límites del intervalo de confianza, y el identificador del modelo usado.

---

8. Aspectos técnicos relevantes

8.1 Arquitectura

El sistema tiene dos componentes independientes comunicados por API REST:

- **Backend** (Python): FastAPI como framework web, scikit-learn para los modelos, SHAP para explicabilidad, pandas para manipulación de datos. Los modelos se almacenan en memoria durante la sesión.

- **Frontend** (TypeScript/React): Material-UI para la interfaz, Recharts y Plotly.js para visualizaciones, Axios como cliente HTTP.

8.2 Validación

El sistema incluye 35 tests automatizados que verifican: - Carga correcta de los 3 datasets - Entrenamiento de modelos con R^2 mínimo aceptable - Consistencia de SHAP (valor base + suma de contribuciones ≈ predicción) - Coherencia de intervalos de confianza (límite inferior ≤ predicción ≤ límite superior) - Persistencia de configuraciones guardadas - Operaciones CRUD completas

8.3 Limitaciones actuales

- Los modelos se entrenan sobre el dataset UCI de concreto, que tiene 1,030 muestras. Para dominios específicos (geopolímeros, concretos de ultra alto rendimiento) se necesitarían datos experimentales propios.
- Los datasets ConcreteXAI y Geopolymer son sintéticos. Están diseñados para demostrar la funcionalidad del sistema, no para reemplazar datos reales.
- Los modelos son de tipo ensemble (Random Forest, Gradient Boosting). No se incluyen redes neuronales ni modelos más complejos, lo cual es una decisión de diseño: para este volumen de datos, los ensembles de árboles ofrecen mejor rendimiento y son compatibles con SHAP de forma nativa.
- La incertidumbre se estima por métodos de ensemble, no por inferencia bayesiana completa. Es una aproximación práctica, no un intervalo de credibilidad riguroso.

---

9. Relevancia y próximos pasos

9.1 Por qué importa

Este trabajo une dos campos que históricamente operan de forma separada:

- La **ciencia de datos** aporta modelos predictivos, cuantificación de incertidumbre y explicabilidad algorítmica.
- La **ingeniería de materiales** aporta el dominio del problema, la interpretación física de los resultados y el criterio para validación experimental.

El resultado es una herramienta que no reemplaza al ingeniero sino que le amplifica la capacidad de exploración. En lugar de probar 10 mezclas en el laboratorio, puede explorar 10,000 computacionalmente, seleccionar las más prometedoras con un criterio informado, y llevar al laboratorio solo las candidatas con mejor predicción y menor incertidumbre.

9.2 Posibles extensiones

- **Datos experimentales propios**: reemplazar los datasets sintéticos por datos generados en el laboratorio del grupo de investigación.
- **Optimización multi-objetivo**: no solo maximizar resistencia sino minimizar costo, huella de carbono o contenido de cemento simultáneamente.

- **Modelos bayesianos**: reemplazar la estimación de incertidumbre por ensemble por Gaussian Process Regression o redes neuronales bayesianas para intervalos de credibilidad más rigurosos.
- **Integración con base de datos de laboratorio**: conectar el sistema con los resultados experimentales reales para un ciclo de retroalimentación datos→ modelo→ exploración→ experimento→ datos.

---

10. Resumen

MDI es una plataforma funcional que permite:

Capacidad	Qué hace
Exploración de datos	Ver estadísticas, correlaciones e histogramas de 3 datasets
Modelado predictivo	Entrenar Random Forest o Gradient Boosting con $R^2 > 0.90$
Predicción con incertidumbre	Dar resistencia estimada + intervalo de confianza al 95%
Explicabilidad SHAP	Descomponer cada predicción en contribuciones por ingrediente
Barrido paramétrico	Variar un ingrediente y ver el efecto en la resistencia
Superficie de respuesta	Variar dos ingredientes y ver un mapa de calor
Comparación de mezclas	Evaluar configuraciones lado a lado con SHAP
Visualización de microestructura	Ver sección transversal 2D del concreto según composición
Gestión de candidatas	Guardar, marcar y exportar mezclas para el laboratorio

El código está escrito en ~3,900 líneas (Python + TypeScript), con 35 tests automatizados pasando, y se ejecuta como aplicación web local sin dependencias externas más allá de las librerías estándar de ciencia de datos.