Práctica de Laboratorio: Machine Learning Aplicado a Ciencia de Materiales

Física Computacional y Machine Learning

1. Información General

- **Asignatura:** Machine Learning
- **Duración:** 1.5 mes
- **Modalidad:** Individual
- **Base de datos:** Materials Project API

2. Descripción del Problema

Los estudiantes desarrollarán modelos de aprendizaje automático supervisado para resolver problemas reales en ciencia de materiales utilizando la base de datos Materials Project, que contiene más de 150,000 materiales computados mediante teoría del funcional de la densidad (DFT).

3. Casos de Estudio

3.1. Caso 1: Clasificación de Materiales Semiconductores

Desarrollar un clasificador binario que identifique materiales semiconductores basándose en sus propiedades compositionales y estructurales.

Entregables:

- Dataset balanceado en formato CSV
- Implementación de al menos 3 algoritmos diferentes
- Análisis de métricas (AUC, precision, recall, F1-score)
- Interpretación de feature importance

3.2. Caso 2: Clasificación de Materiales Fotovoltaicos

Crear un modelo que clasifique materiales como aptos o no aptos para aplicaciones fotovoltaicas.

Entregables:

- Dataset procesado con ingeniería de características
- Comparación de algoritmos de clasificación

- Matriz de confusión y curvas ROC
- Análisis de errores y casos límite

3.3. Caso 3: Predicción de Estabilidad Termodinámica

Desarrollar un clasificador que prediga si un material es termodinámicamente estable y por tanto sintetizable experimentalmente.

Entregables:

- Dataset con técnicas de balanceamiento aplicadas
- Validación cruzada estratificada
- Optimización de hiperparámetros
- Evaluación en conjunto de prueba independiente

3.4. Caso 4: Predicción de Band Gap (Regresión)

Implementar modelos de regresión para predecir el valor numérico del band gap de materiales semiconductores.

Entregables:

- Dataset de regresión con variables continuas
- Implementación de modelos de regresión (lineal, tree-based, ensemble)
- Métricas de regresión (MSE, MAE, R^2)
- Análisis de residuos y validación del modelo

4. Diccionario de Variables

A continuación se presenta la descripción completa de todas las variables disponibles en el dataset:

Variable	Tipo	Descripción
`material_id`	String	Identificador único del material en Materials Project (formato: mp-XXXXX)
`formula`	String	Fórmula química reducida del material (ej: SiO2, GaAs)
`band_gap`	Float	Band gap del material en eV calculado mediante DFT
`formation_energy_per`	Float	Energía de formación por átomo en eV/atom
`energy_above_hull`	Float	Energía sobre el hull de estabilidad en eV/atom
`density`	Float	Densidad del material en g/cm^3
`nsites`	Integer	Número total de sitios atómicos en la celda unitaria
`volume`	Float	Volumen de la celda unitaria en Ų
`volume_per_atom`	Float	Volumen por átomo en Ų/atom

Variable	Tipo	Descripción
crystal_system	String	Sistema cristalino (cubic, tetragonal, orthorhombic, hexagonal, trigonal, monoclinic, triclinic)
spacegroup_number	Integer	Número del grupo espacial (1-230)
spacegroup_symbol	String	Símbolo del grupo espacial en notación internacional
n_elements	Integer	Número de elementos químicos únicos en la composición
avg_atomic_mass	Float	Masa atómica promedio ponderada en uma
min_atomic_mass	Float	Masa atómica mínima entre todos los elementos
max_atomic_mass	Float	Masa atómica máxima entre todos los elementos
avg_electronegativit	Float	Electronegatividad promedio ponderada (escala de Pauling)
min_electronegativit	Float	Electronegatividad mínima entre todos los elementos
max_electronegativit	Float	Electronegatividad máxima entre todos los elementos
electronegativity_ra	Float	Diferencia entre electronegatividad máxima y mínima
avg_atomic_radius	Float	Radio atómico promedio ponderado en pm
avg_ionic_radius	Float	Radio iónico promedio ponderado en pm
has_metal	Boolean	Indica si la composición contiene al menos un elemento metálico (1/0)
has_metalloid	Boolean	Indica si la composición contiene al menos un metaloide (1/0)
has_nonmetal	Boolean	Indica si la composición contiene al menos un no-metal (1/0)
n_metals	Integer	Número de elementos metálicos en la composición
n_nonmetals	Integer	Número de elementos no-metálicos en la composición
n_metalloids	Integer	Número de metaloides en la composición
is_binary	Boolean	Indica si el material tiene exactamente 2 elementos (1/0)
is_ternary	Boolean	Indica si el material tiene exactamente 3 elementos (1/0)
is_quaternary	Boolean	Indica si el material tiene exactamente 4 elementos (1/0)

4.1. Fracciones Atómicas de Elementos

Para cada elemento importante en semiconductores y materiales avanzados, se incluye su fracción atómica:

Variable	Tipo	Descripción
frac_H	Float	Fracción atómica de Hidrógeno (0.0-1.0)

Variable	Tipo	Descripción
frac_Li	Float	Fracción atómica de Litio (0.0-1.0)
frac_C	Float	Fracción atómica de Carbono (0.0-1.0)
frac_N	Float	Fracción atómica de Nitrógeno (0.0-1.0)
frac_O	Float	Fracción atómica de Oxígeno (0.0-1.0)
frac_F	Float	Fracción atómica de Flúor (0.0-1.0)
frac_Na	Float	Fracción atómica de Sodio (0.0-1.0)
frac_Mg	Float	Fracción atómica de Magnesio (0.0-1.0)
frac_Al	Float	Fracción atómica de Aluminio (0.0-1.0)
frac_Si	Float	Fracción atómica de Silicio (0.0-1.0)
frac_P	Float	Fracción atómica de Fósforo (0.0-1.0)
frac_S	Float	Fracción atómica de Azufre (0.0-1.0)
frac_Cl	Float	Fracción atómica de Cloro (0.0-1.0)
frac_K	Float	Fracción atómica de Potasio (0.0-1.0)
frac_Ca	Float	Fracción atómica de Calcio (0.0-1.0)
frac_Ti	Float	Fracción atómica de Titanio (0.0-1.0)
frac_V	Float	Fracción atómica de Vanadio (0.0-1.0)
frac_Cr	Float	Fracción atómica de Cromo (0.0-1.0)
frac_Mn	Float	Fracción atómica de Manganeso (0.0-1.0)
frac_Fe	Float	Fracción atómica de Hierro (0.0-1.0)
frac_Co	Float	Fracción atómica de Cobalto (0.0-1.0)
frac_Ni	Float	Fracción atómica de Níquel (0.0-1.0)
frac_Cu	Float	Fracción atómica de Cobre (0.0-1.0)
frac_Zn	Float	Fracción atómica de Zinc (0.0-1.0)
frac_Ga	Float	Fracción atómica de Galio (0.0-1.0)
frac_Ge	Float	Fracción atómica de Germanio (0.0-1.0)
frac_As	Float	Fracción atómica de Arsénico (0.0-1.0)
frac_Se	Float	Fracción atómica de Selenio (0.0-1.0)
frac_Br	Float	Fracción atómica de Bromo (0.0-1.0)
frac_Sr	Float	Fracción atómica de Estroncio (0.0-1.0)
frac_Y	Float	Fracción atómica de Itrio (0.0-1.0)
frac_Zr	Float	Fracción atómica de Circonio (0.0-1.0)
frac_Nb	Float	Fracción atómica de Niobio (0.0-1.0)
frac_Mo	Float	Fracción atómica de Molibdeno (0.0-1.0)
frac_Ru	Float	Fracción atómica de Rutenio (0.0-1.0)
frac_Rh	Float	Fracción atómica de Rodio (0.0-1.0)
frac_Pd	Float	Fracción atómica de Paladio (0.0-1.0)
frac_Ag	Float	Fracción atómica de Plata (0.0-1.0)
frac_Cd	Float	Fracción atómica de Cadmio (0.0-1.0)
frac_In	Float	Fracción atómica de Indio (0.0-1.0)
frac_Sn	Float	Fracción atómica de Estaño (0.0-1.0)
frac_Sb	Float	Fracción atómica de Antimonio (0.0-1.0)
frac_Te	Float	Fracción atómica de Telurio (0.0-1.0)
frac_I	Float	Fracción atómica de Yodo (0.0-1.0)
frac_Ba	Float	Fracción atómica de Bario (0.0-1.0)
frac_La	Float	Fracción atómica de Lantano (0.0-1.0)
frac_Ce	Float	Fracción atómica de Cerio (0.0-1.0)
frac_Hf	Float	Fracción atómica de Hafnio (0.0-1.0)
frac_Ta	Float	Fracción atómica de Tantalio (0.0-1.0)

Variable	Tipo	Descripción
frac_W	Float	Fracción atómica de Tungsteno (0.0-1.0)
frac_Re	Float	Fracción atómica de Renio (0.0-1.0)
frac_Os	Float	Fracción atómica de Osmio (0.0-1.0)
frac_Ir	Float	Fracción atómica de Iridio (0.0-1.0)
frac_Pt	Float	Fracción atómica de Platino (0.0-1.0)
frac_Au	Float	Fracción atómica de Oro (0.0-1.0)
frac_Hg	Float	Fracción atómica de Mercurio (0.0-1.0)
frac_Tl	Float	Fracción atómica de Talio (0.0-1.0)
frac_Pb	Float	Fracción atómica de Plomo (0.0-1.0)
frac_Bi	Float	Fracción atómica de Bismuto (0.0-1.0)

4.2. Variables Target (Etiquetas)

Variable	Tipo	Descripción
is_semiconductor	Boolean	Etiqueta binaria: 1 si es semiconductor (0.1 <band_gap <4.0 eV), 0 en caso contrario
is_photovoltaic	Boolean	Etiqueta binaria: 1 si es apto para aplicaciones fotovoltaicas ($0{,}8 \leq$ band_gap $\leq 2{,}2$ eV), 0 en caso contrario
is_stable	Boolean	Etiqueta binaria: 1 si es termodinámicamente estable (energy_above_hull <0.1 eV), 0 en caso contrario

5. Especificaciones Técnicas

5.1. Datos y Características

- Extraer mínimo 1000 materiales por caso
- Generar al menos 50 características por material
- Incluir propiedades compositionales, estructurales y termodinámicas
- Documentar todas las fuentes de datos y transformaciones

5.2. Modelos Requeridos

Cada caso debe incluir al menos:

- Un modelo baseline (regresión logística o lineal)
- Un modelo basado en árboles (Random Forest o Gradient Boosting)
- Un modelo de su elección justificando la selección

5.3. Análisis y Visualización

- Análisis exploratorio de datos completo
- Visualizaciones de distribuciones y correlaciones
- Gráficos de rendimiento de modelos
- Interpretabilidad y feature importance

6. Entregables Finales

1. **Código Python completo** con documentación
2. **Artículo** (máximo 15 páginas) incluyendo:
   - Metodología de extracción y procesamiento de datos
   - Justificación de selección de características
   - Comparación de algoritmos y métricas
   - Interpretación de resultados desde perspectiva física
   - Limitaciones y trabajo futuro
3. **Presentación oral** (10 minutos + 5 minutos preguntas)

7. Criterios de Evaluación

- **Calidad técnica del código** (25 %)
- **Interpretación física de resultados** (25 %)
- **Originalidad y profundidad del análisis** (25 %)
- **Presentación** (25 %)

8. Recursos Disponibles

- API key de Materials Project (proporcionada por el instructor)
- Tutoriales de pymatgen y mp-api
- Sesiones de consulta semanales
- Cluster de cómputo departamental

9. Fechas Importantes

- **Semana 1:** Extracción y exploración inicial de datos
- **Semana 2:** Ingeniería de características y modelos baseline
- **Semana 3:** Optimización de modelos y análisis avanzado
- **Semana 4:** GPU laboratorio
- **Entrega final:** [Fecha específica según calendario académico]

10. Nota Importante

Este proyecto requiere comprensión tanto de conceptos de machine learning como de principios básicos de ciencia de materiales. Se espera que los estudiantes consulten literatura científica relevante para interpretar correctamente sus resultados.

Total de variables disponibles: 87 características + 3 variables target