



Ingeniería de características basada en ontologías para mejorar el rendimiento de modelos de machine learning

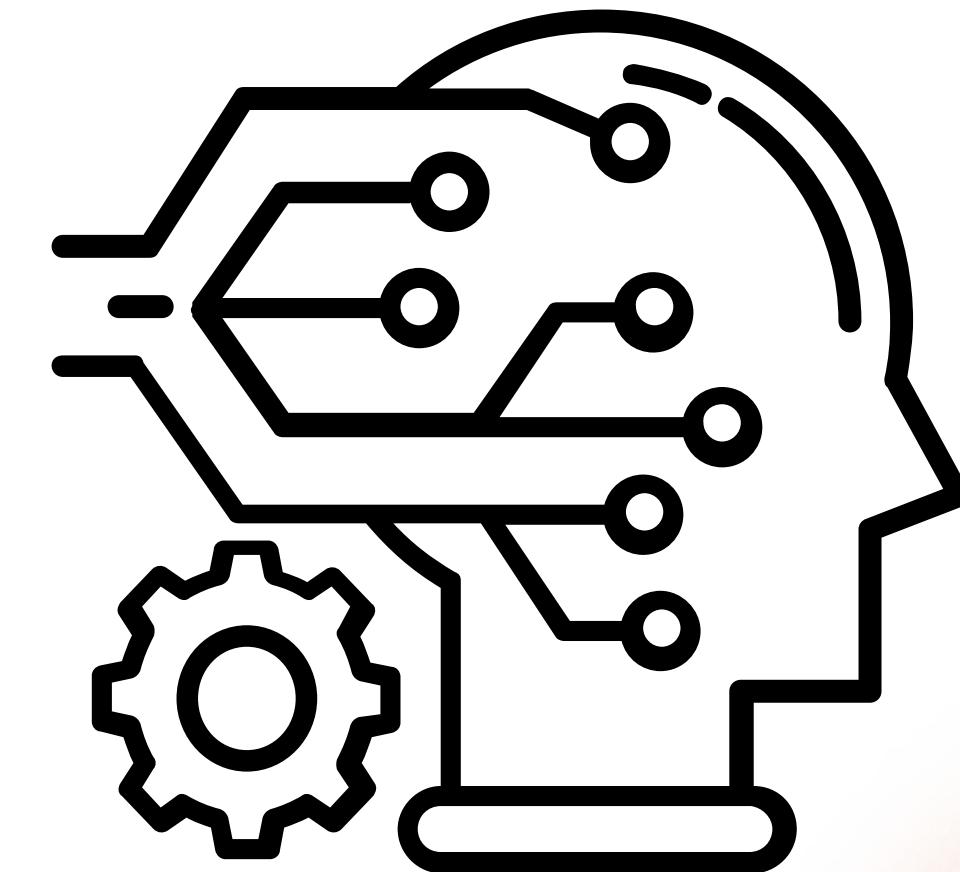
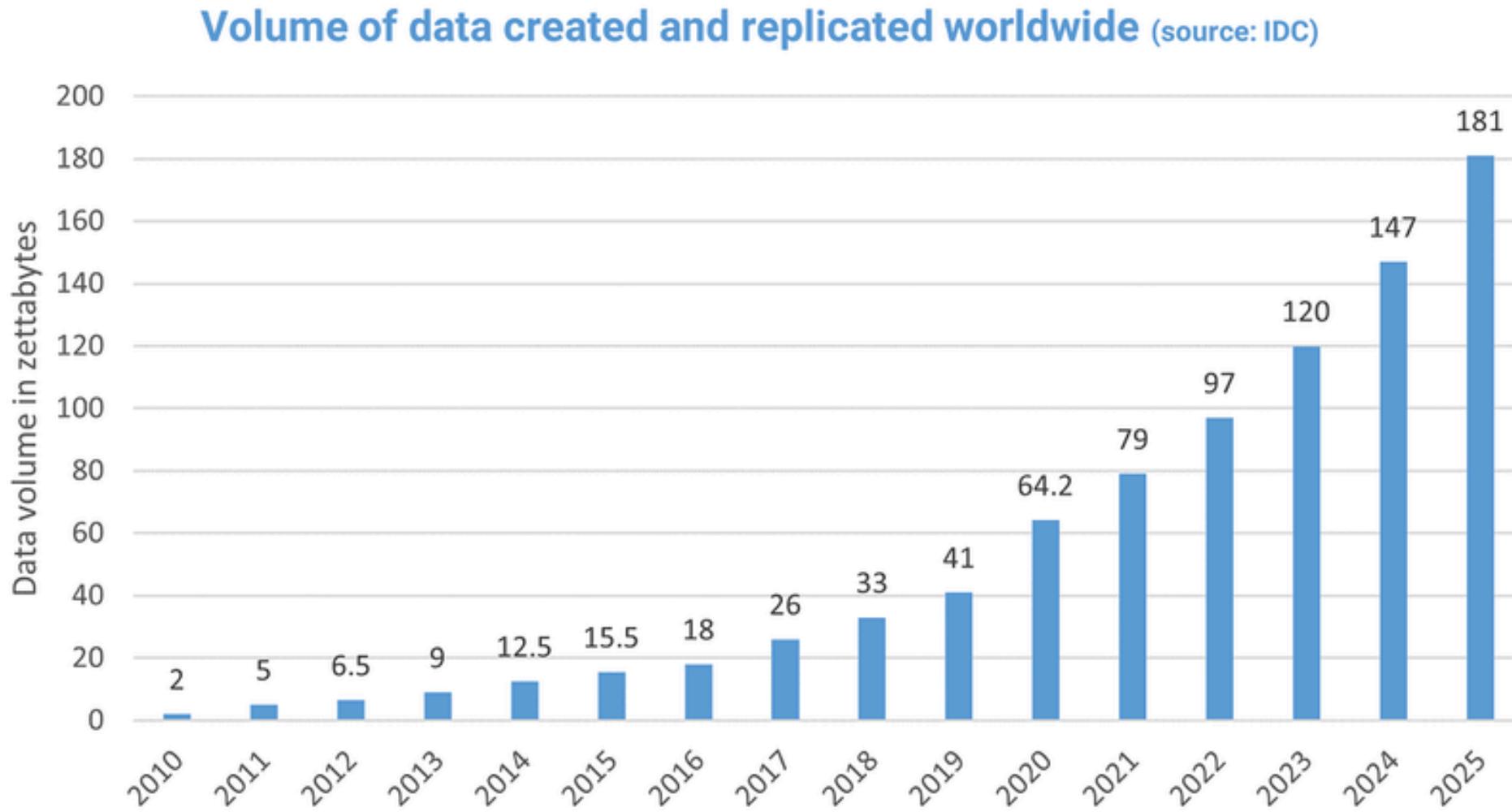
Daniela Jiménez

Tutor: Ricardo Flores, Ph.D

Universidad San Francisco de Quito

Motivación

Contexto General



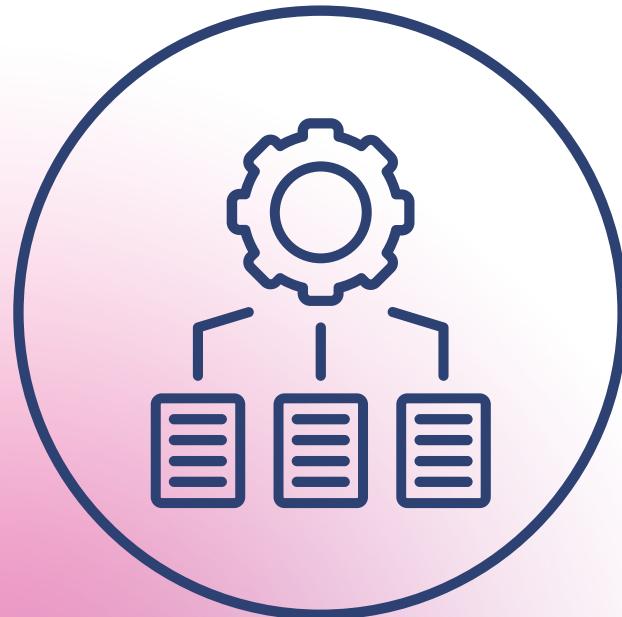
- Vivimos en una era de sobrecarga de información
- Los datos se generan a una velocidad sin precedentes
- Más datos ≠ más conocimiento útil
- *Machine Learning* necesita datos de calidad para funcionar
- Éxito del modelo no depende solo del algoritmo
- **Clave:** preparación y representación de datos

Motivación

Ingeniería de características (Feature engineering)

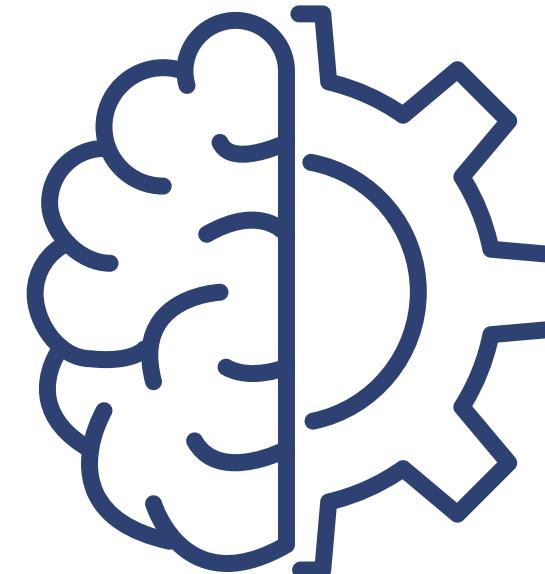
Qué es?

Proceso de seleccionar, transformar y organizar atributos. Además, define qué ve el modelo y cómo lo interpreta.



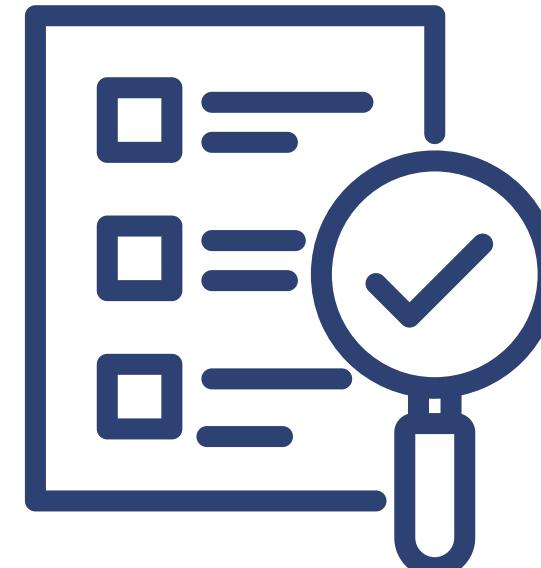
Importancia

Una de las etapas más críticas en ML que impacta directamente en la calidad del aprendizaje



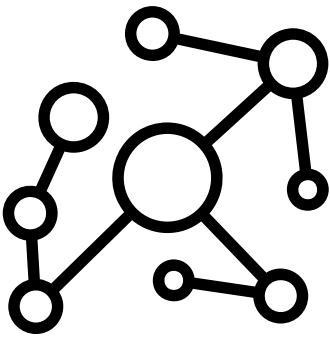
Tendencia actual

- Crece el interés por métodos estructurados e inteligentes
- Mejores representaciones = Mejores resultados



Ontologías para ingeniería de características

Ontología



Representación formal y jerárquica del conocimiento.

No solo términos → también relaciones, clases y propiedades.

Componentes Clave

- **Clases:** Producto, Cliente, Transacción...
- **Relaciones:** Cliente realiza Transacción
- **Jerarquías:** "Tarjeta de crédito" \subseteq "Método de pago"

Por qué es útil?

Ayuda a estructurar datos complejos para facilitar el entendimiento y la reutilización del conocimiento. Mejora la calidad de las features para ML.

Tecnologías usadas

- OWL (Web Ontology Language)
- Protégé para modelar y visualizar



Ontologías para ingeniería de características

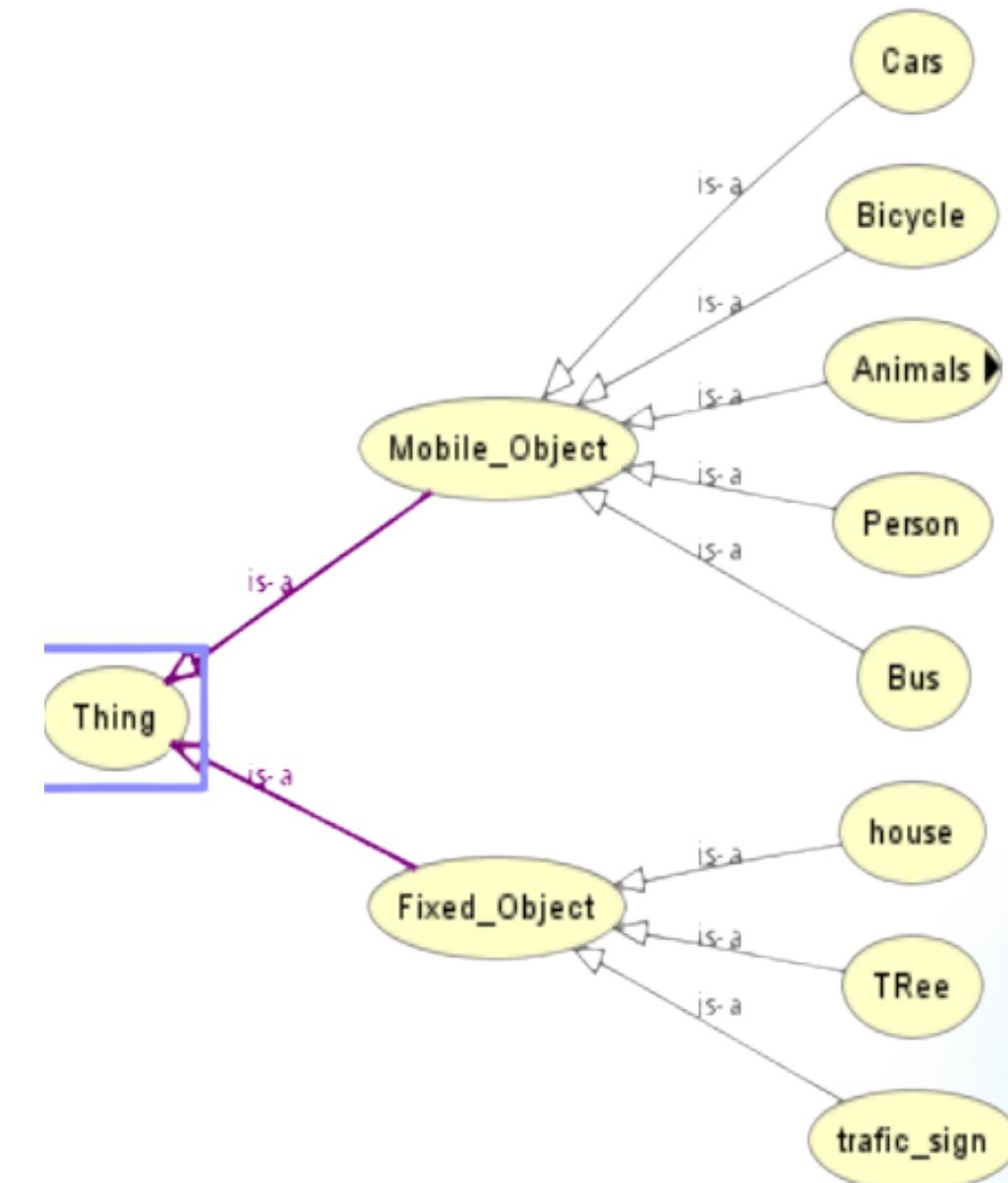
Ontologías + Machine Learning

Dónde encajan?

Apoyan la ingeniería de características, etapa clave en ML, porque sin buena representación, incluso el mejor algoritmo puede fallar.

Qué aportan?

1. Guiado semántico
2. Agrupación inteligente
3. Reducción de errores
4. Mejor interpretación



PLUS:

- Las ontologías actúan como una capa de conocimiento previa al modelo.

Estado del arte: Investigaciones previas

Faust et al. (2019): Ontologías + Deep Learning en imágenes médicas

- Estandarizaron morfologías de tumores cerebrales.
- Un modelo más estructurado y preciso dio un mejor desempeño.

Sahoo et al. (2022): Epilepsia y datos clínicos heterogéneos

- Ontología para integrar registros médicos
- Mejor clasificación multietiqueta y menor tiempo de entrenamiento

Siddiqui et al. (2019): Minería de opiniones y análisis de sentimientos

- Agrupación semántica de conceptos
- Features más relevantes dieron mayor precisión

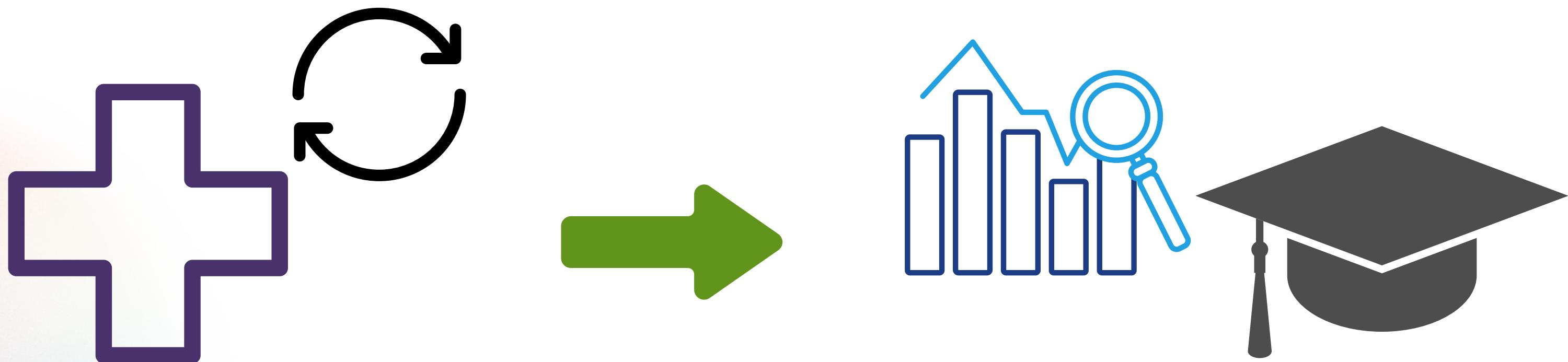
Problema

1. Uso limitado de ontologías en ML para otros dominios

- Han mostrado gran potencial
- Uso sigue siendo limitado y focalizado
- Principalmente aplicadas en áreas como medicina / biomedicina.

2. Oportunidad desaprovechada

- Dominios complejos como la educación aún poco explorados
- Falta de aplicaciones que aprovechen su capacidad de estructurar y enriquecer datos



Problema

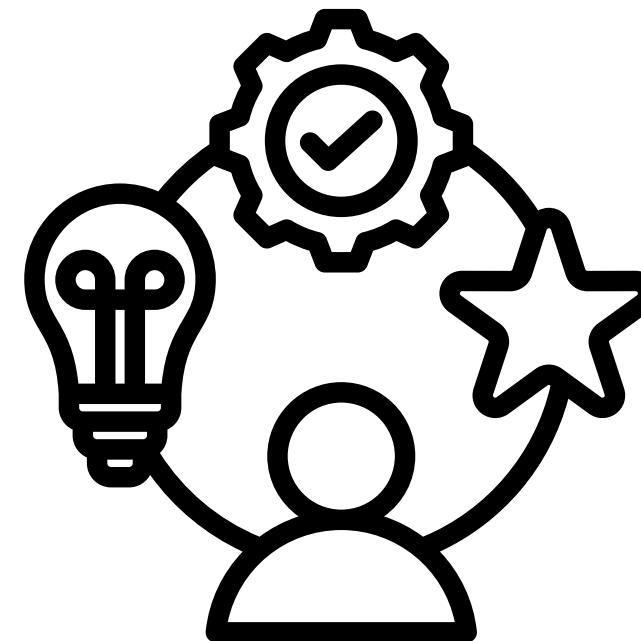
3. Vacío en el dominio educativo

- Volumen elevado de datos: notas, asistencia, participación, pero con análisis basados en estadística genérica.



4. Limitaciones actuales

- Ausencia de capa semántica y conocimiento experto para contextualizar datos
- *Enfoque común:* Predicción de riesgo de diserción



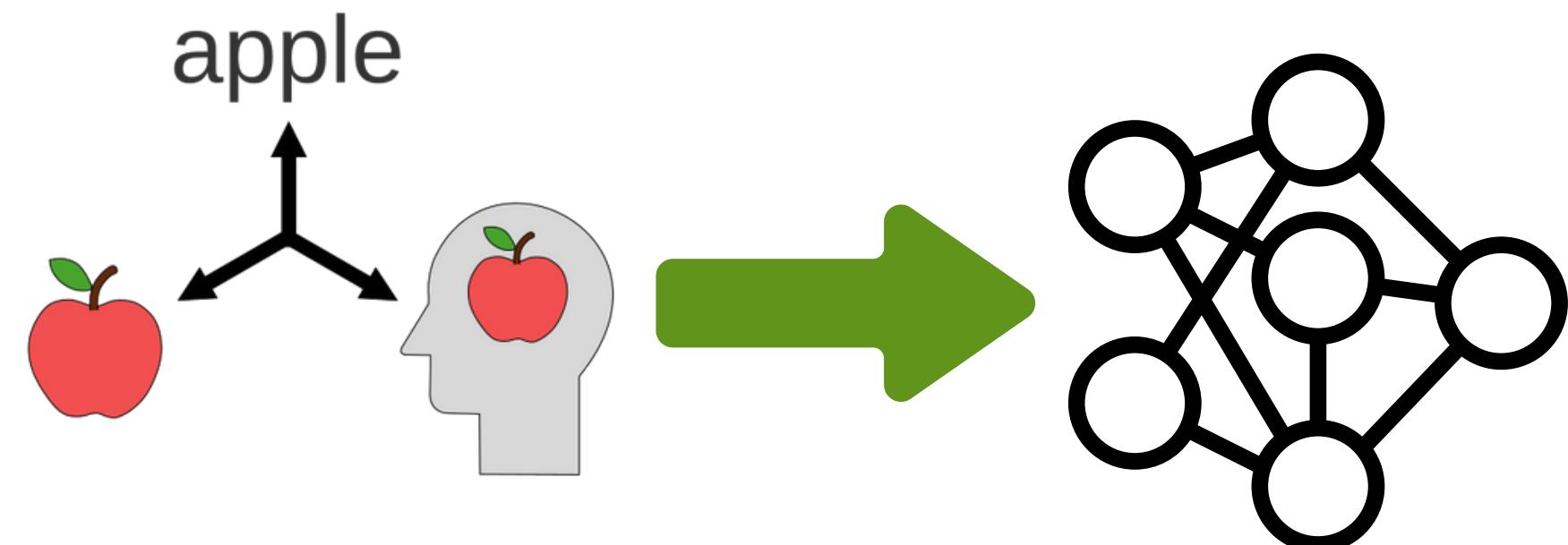
Solución

Propuesta

Desarrollar un prototipo que incorpore una ontología del ámbito en el proceso de ingeniería de características.

¿Cómo se implementa?

1. Estandarización semántica
2. Mapeo semántico guiado
3. Enriquecimiento de features



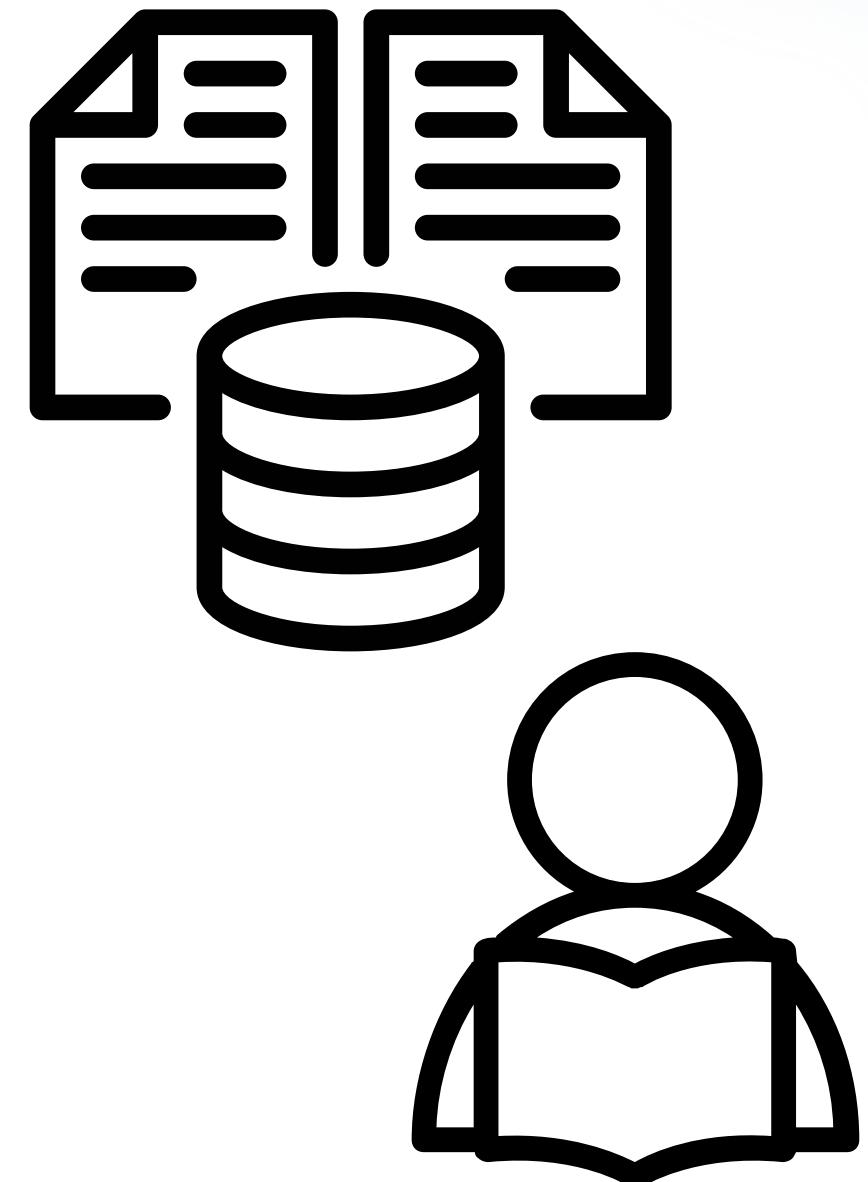
Dataset: Student Performance

Overview

- *Fuente:* UCI Machine Learning Repository
- *Contexto:* Estudiantes de secundaria en Portugal (Matemáticas)
- *Tamaño:* 649 registros

Atributos (33 en 6 dimensiones)

1. Académicos
2. Sociodemográficos
3. Familiares
4. Hábitos de Estudio:
5. Sociales & Conductuales
6. Infraestructura y acceso



Dataset: Student Performance

¿Por qué este dataset?

- *Complejidad semántica*: combina variables académicas, familiares, emocionales y más
- Versatilidad experimental.

Aplicación de la Ontología

Ontología personalizada creada para el dominio educativo

Mapeo semántico: atributo → clase conceptual

Dominios Conceptuales Utilizados

AcademicPerformance

StudyHabits

FamilyContext

HealthAndWellBeing

SocialBehaviour

InfrastructureAccess

ParentsOccupation

Demographics



Metodología: Diseño y desarrollo de la ontología

Análisis y agrupación

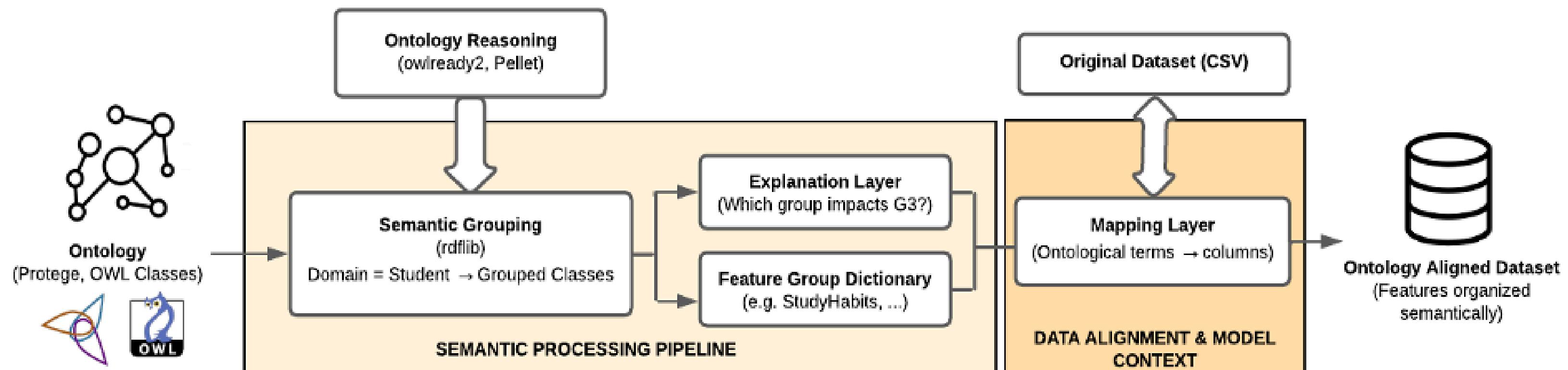
- *Dataset*: identificación de conceptos y atributos clave.
- Variables organizadas en dominios semánticos: académico, estudio, salud, social, familiar...

Modelado en OWL

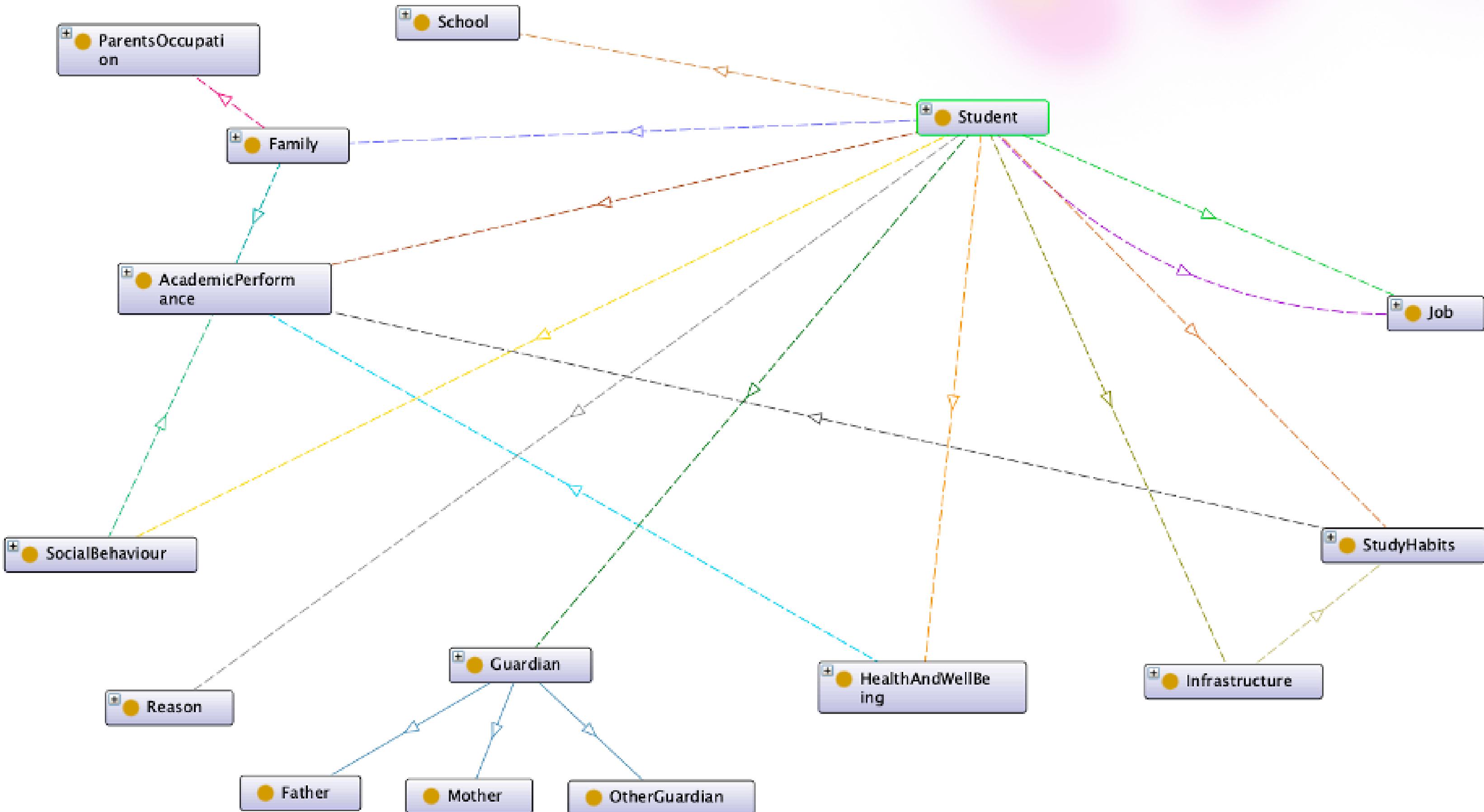
- Definición de clases (Student, Family, AcademicPerformance, ...)
- Establecimiento de propiedades y jerarquías entre clases

Resultado

Datos planos → representación estructurada y enriquecida del conocimiento educativo



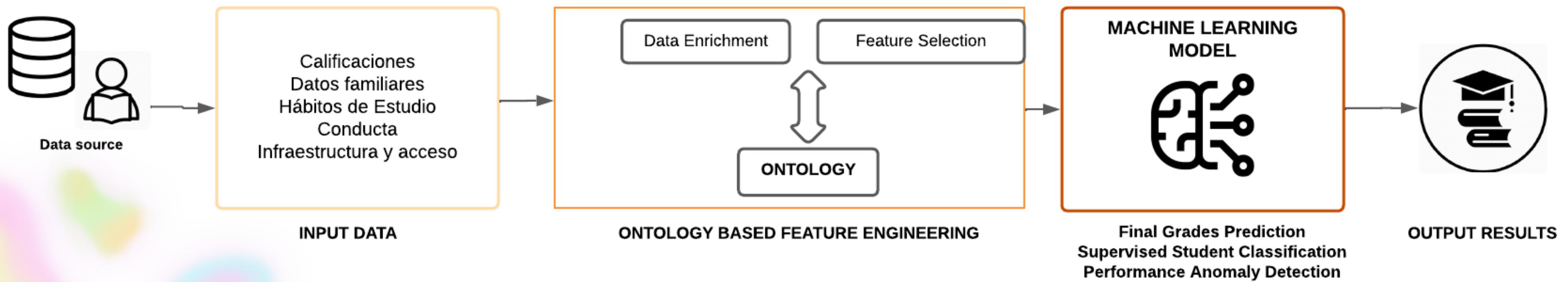
Metodología: Desarrollo del prototipo ontológico



Metodología: Integración en el pipeline de ML

Con la ontología definida:

- **Comparación de modelos tradicionales vs. enriquecidos con ontología**
- **Semantic feature grouping:** Agrupación de variables según sus clases ontológicas
- **Implementación en Python:** Uso de owlready2 y rdflib para el mapeo de características
- **Evaluación de impacto**



Metodología: Configuración y pruebas de los experimentos y Análisis de resultados

Configuración de Experimentos

3 tareas:

Regresión (predicción de nota final)

Clasificación (alto/medio/bajo rendimiento)

Detección de anomalías (No supervisada)

2 modelos base:

Features tradicionales (seleccionadas manuales y por correlación estadística)

Modelo Ontológico:

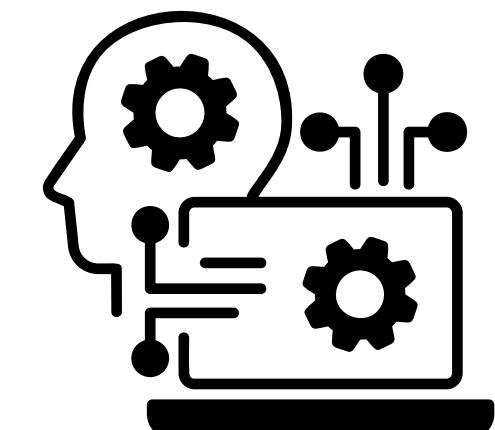
Features semánticos (agrupados por ontología)

Evaluación y Métricas

- *Regresión*: MSE, R²
- *Clasificación*: Accuracy, Precision, Recall, F1-score
- Anomalías: Accuracy, Precision, Recall, F1-score, AUC-ROC, AUC-PR

Comparaciones clave:

- Mejora de rendimiento numérico en un caso
- Robustez sin fine-tuning



Casos experimentales y resultados

Caso 1: Predicción de calificaciones finales (regresión)

Modelos evaluados

- Base: 8 features manuales (G1, G2, studytime, failures, absences, goout, Dalc, Walc)
- Top 10: 10 variables mejor correlacionadas con G3
- Ontológico: 26 features agrupadas semánticamente

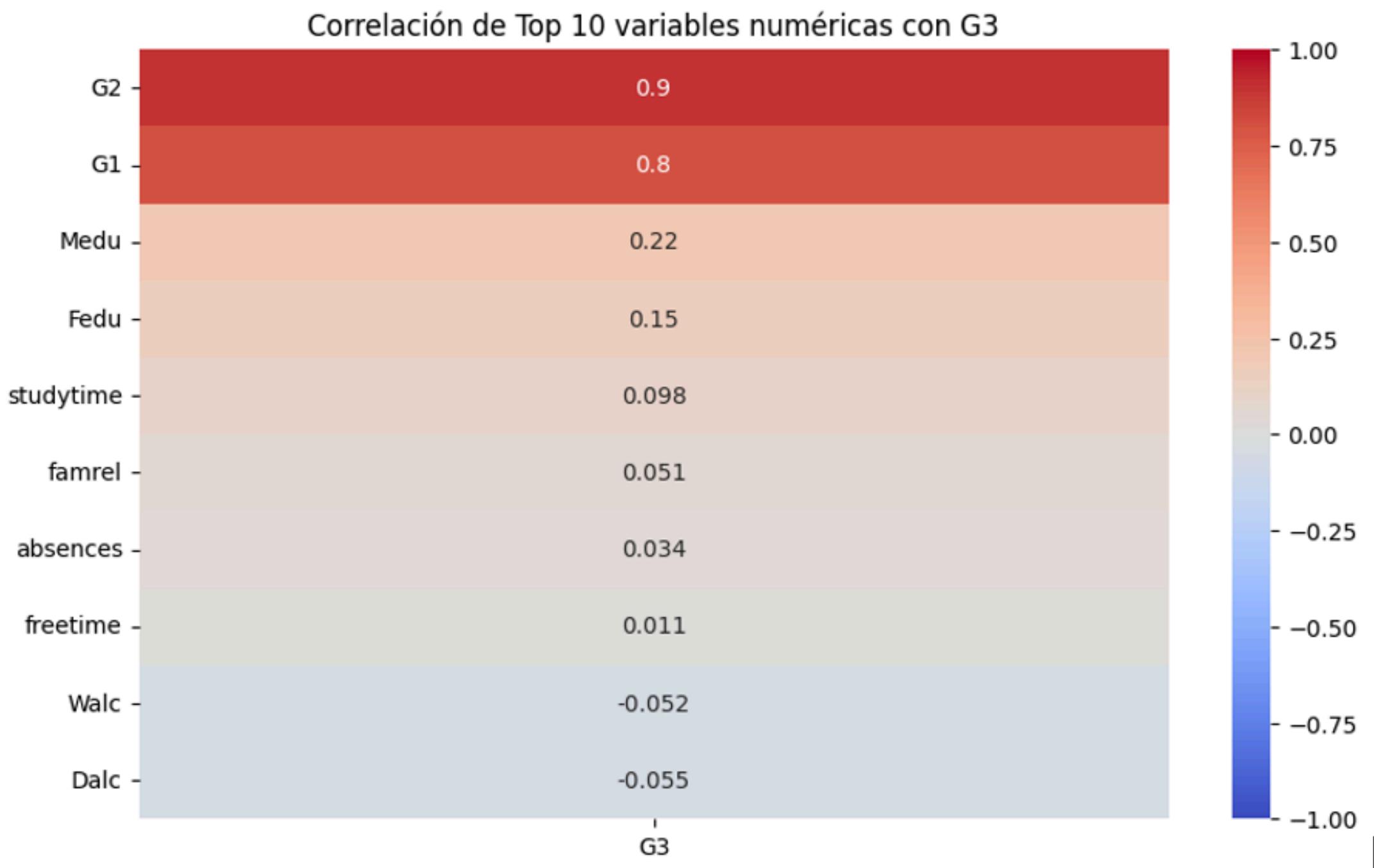
Resultados (RandomForestRegressor)

| Modelo | MSE | R ² |
|---------------------------------|------|----------------|
| Base (8 features) | 3.20 | 0.84 |
| Top 10 features correlacionadas | 2.92 | 0.86 |
| Ontológico | 3.54 | 0.83 |

Conclusión

- El modelo ontológico rinde de forma competitiva
- Mayor interpretabilidad y estructura semántica para futuras generalizaciones

Figura 6. 10 variables numéricas más correlacionadas con la calificación final (G3).



Caso 2: Clasificación de estudiantes por desempeño académico

Modelos evaluados

- Base: 8 features manuales (G1, G2, studytime, failures, absences, goout, Dalc, Walc)
- Top 15: 15 variables mejor correlacionadas con G3
- Ontológico: 16 features agrupadas semánticamente

Resultados (RandomForestClassifier)

| Modelo | F1 (Alto) | F1 (Medio) | F1 (Bajo) | Accuracy |
|---------------------------------|-----------|------------|-----------|----------|
| Base (8 features) | 0.94 | 0.88 | 0.94 | 0.92 |
| Top 15 features correlacionadas | 0.97 | 0.91 | 0.95 | 0.94 |
| Ontológico | 0.97 | 0.91 | 0.95 | 0.939 |

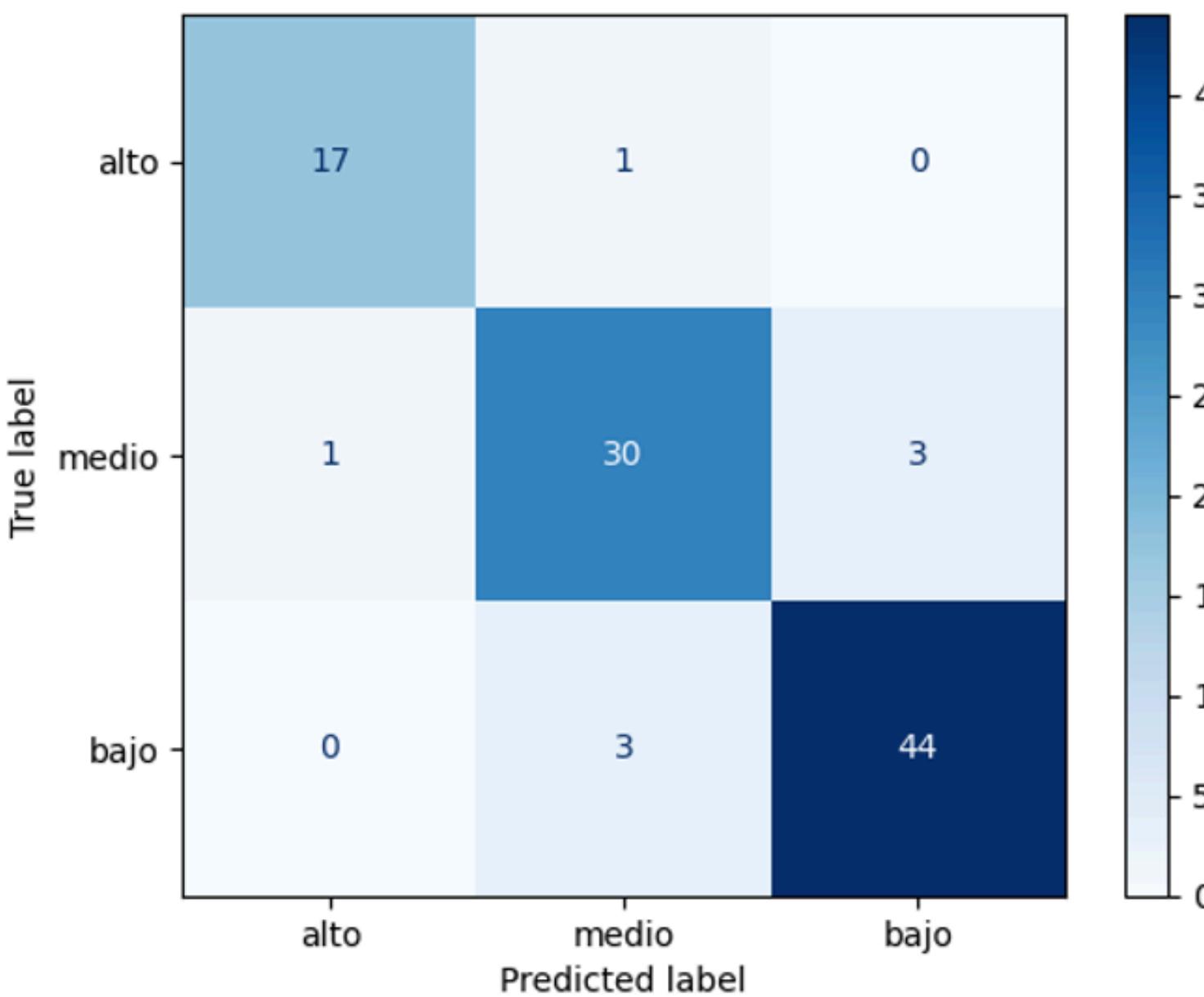
Categorías

Bajo (≤ 10) | Medio (11–14) | Alto (≥ 15)

Ventajas del modelo ontológico

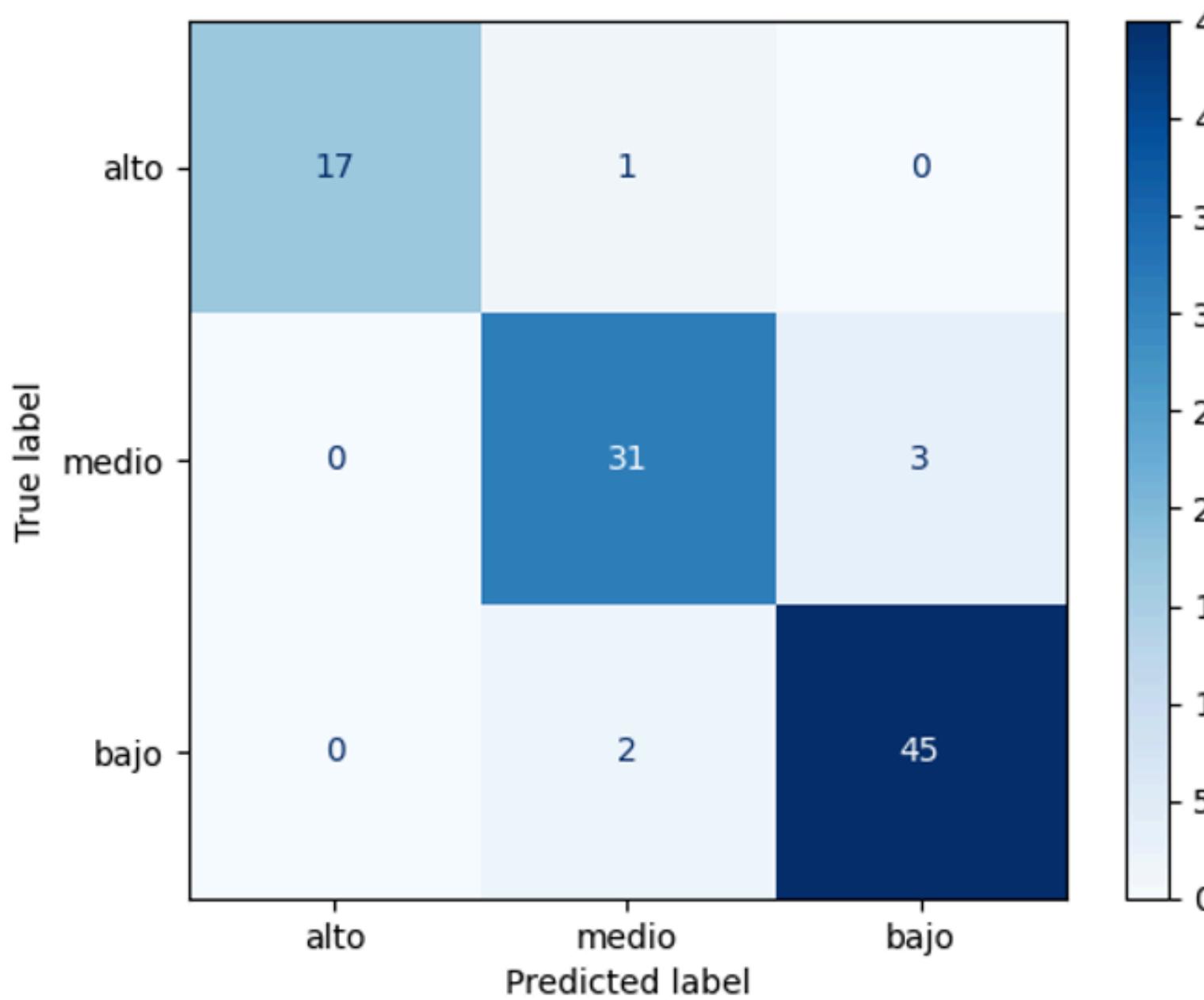
- Rendimiento igual o superior
- Mejor balance entre clases
- Trazabilidad semántica de los atributos.

Figura 8. Matriz de confusión para modelo base con selección manual de features.



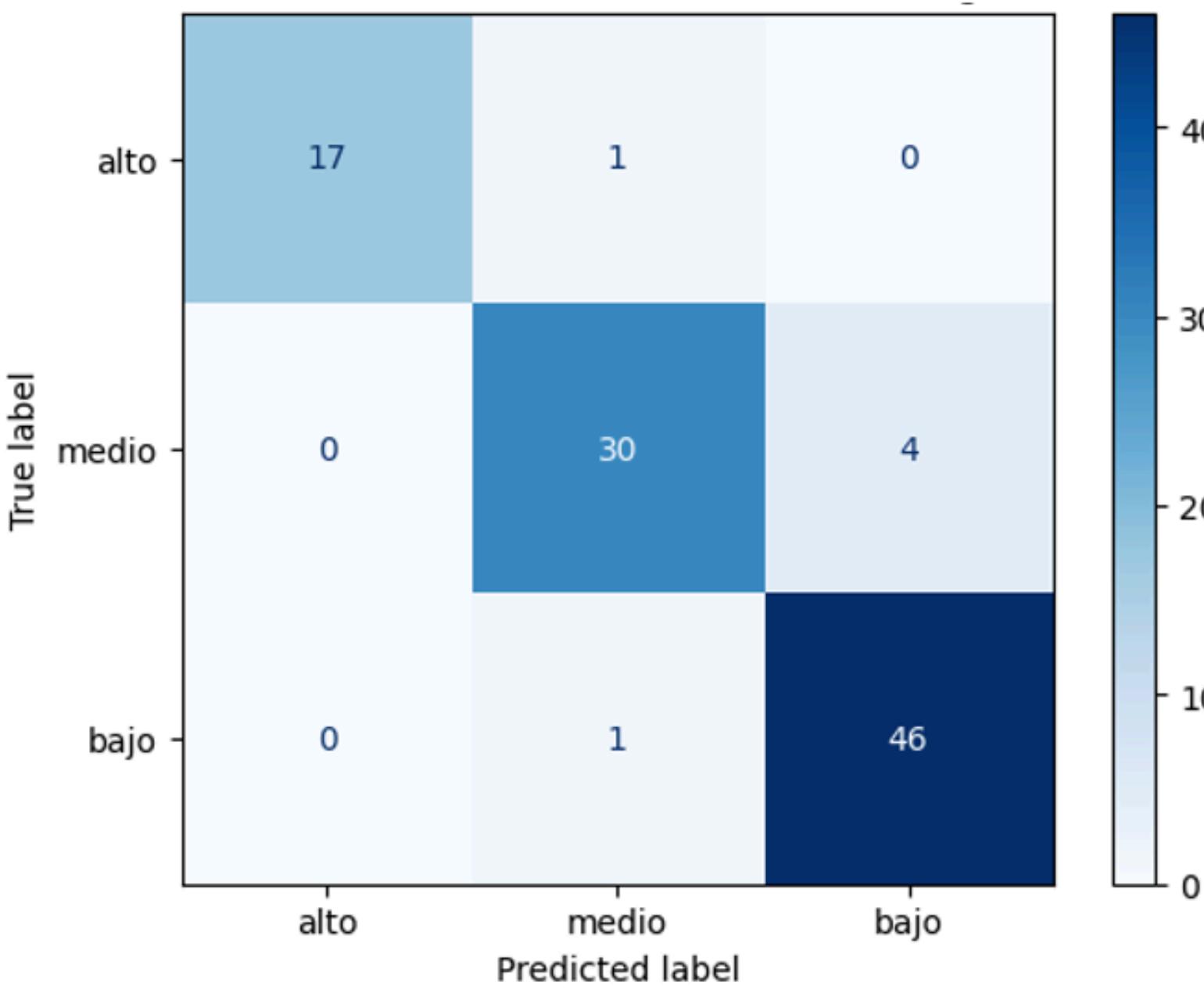
Nota. Se observa una confusión moderada entre las clases “medio” y “bajo”, lo cual indica que estas categorías comparten patrones similares cuando se utilizan variables seleccionadas por criterio experto.

Figura 9. Matriz de confusión del modelo con top 15 features correlacionadas.



Nota. La precisión mejora en la clase "medio", reduciendo los errores de clasificación observados en el modelo de selección manual de features, lo que refleja el impacto positivo de la selección automática por correlación.

Figura 10. Matriz de confusión del modelo ontológico.



Nota. El modelo mantiene un rendimiento casi equilibrado entre clases, con resultados competitivos y una distribución de errores comparable a los modelos tradicionales.

Caso 3: Detección de anomalías en el rendimiento educativo

Modelos evaluados:

1. Base (8 features + GridSearch)
2. Top 10 (features con mayor correlación + GridSearch)
3. Ontológico (Sin ajuste fino, 15 features agrupadas semánticamente)

Objetivo: identificar estudiantes con >2 materias reprobadas.

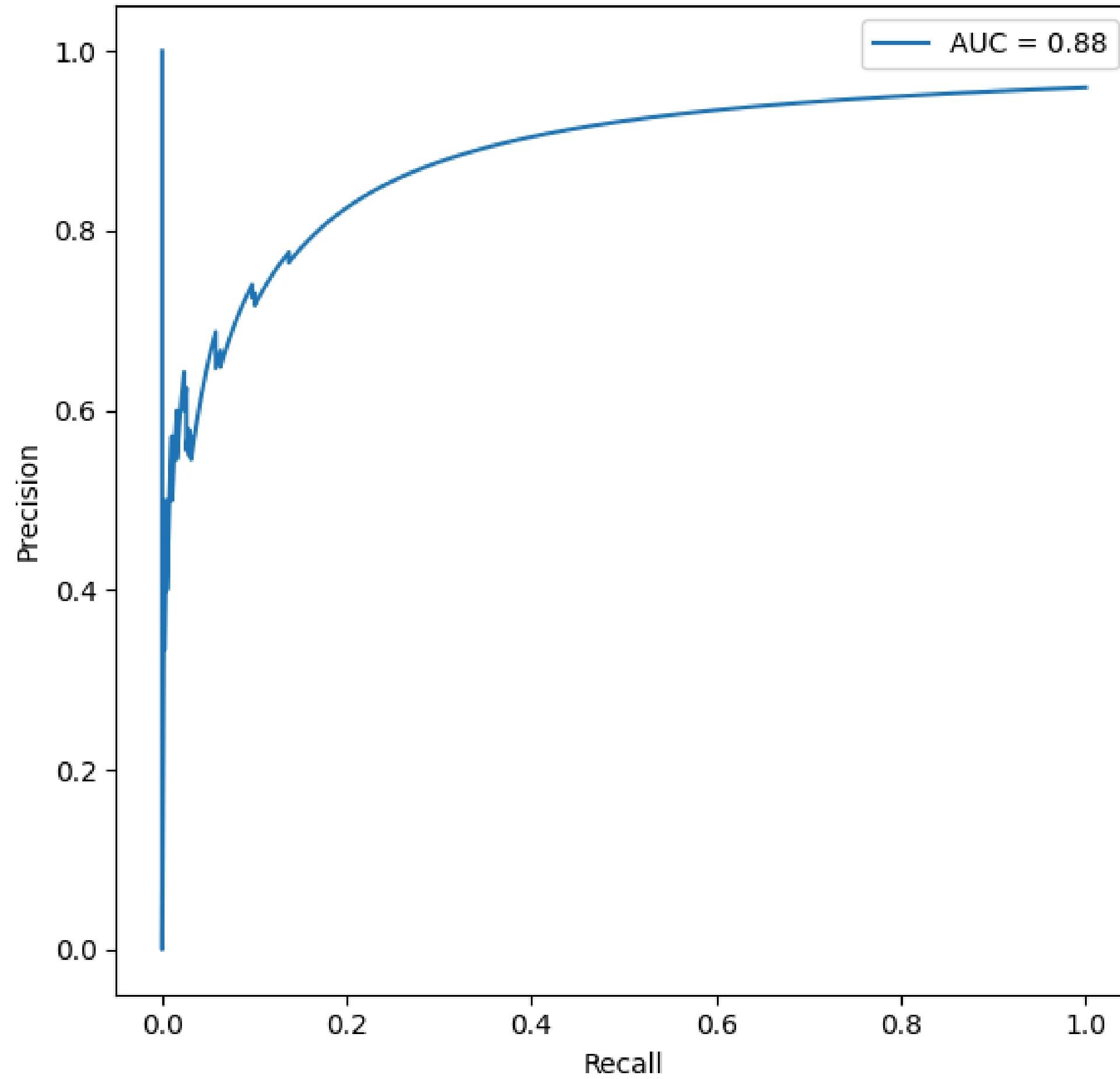
Resultados (IsolationForest)

| Modelo | F1-Score (-1) | Recall (-1) | Precision (-1) | Accuracy | AUC ROC | AUC PR |
|---------------------------------|------------------|----------------|-------------------|----------|---------|--------|
| Base (8 features) | 0.39 | 0.94 | 0.25 | 0.88 | 0.91 | 0.88 |
| Top 10 features correlacionadas | 0.33 | 0.67 | 0.22 | 0.90 | 0.78 | 0.90 |
| Ontológico | 0.20 | 0.33 | 0.14 | 0.90 | 0.63 | 0.93 |

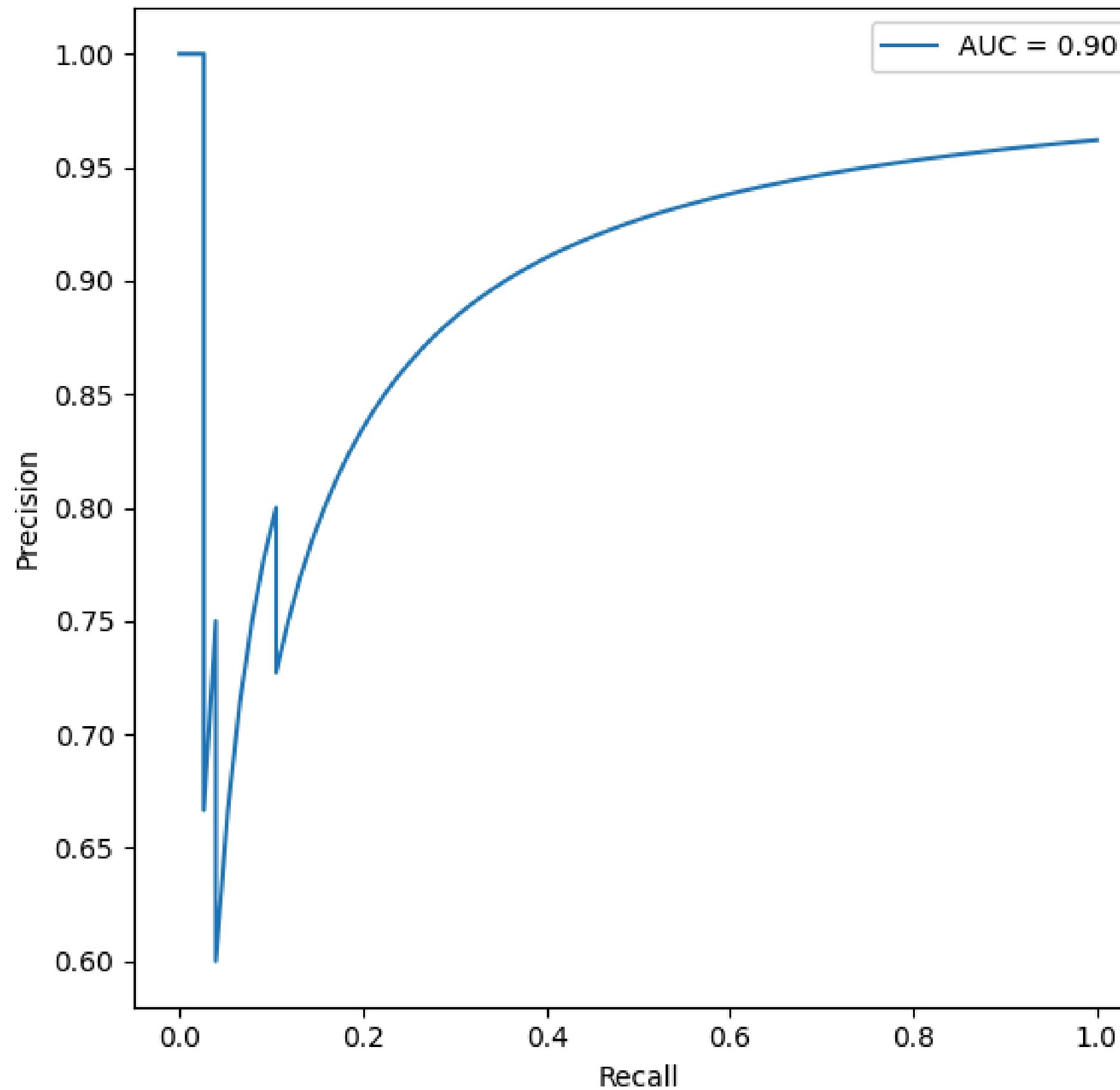
Conclusión:

- El modelo ontológico captura patrones relevantes sin overfitting
- Su regularización semántica aporta robustez y generalización

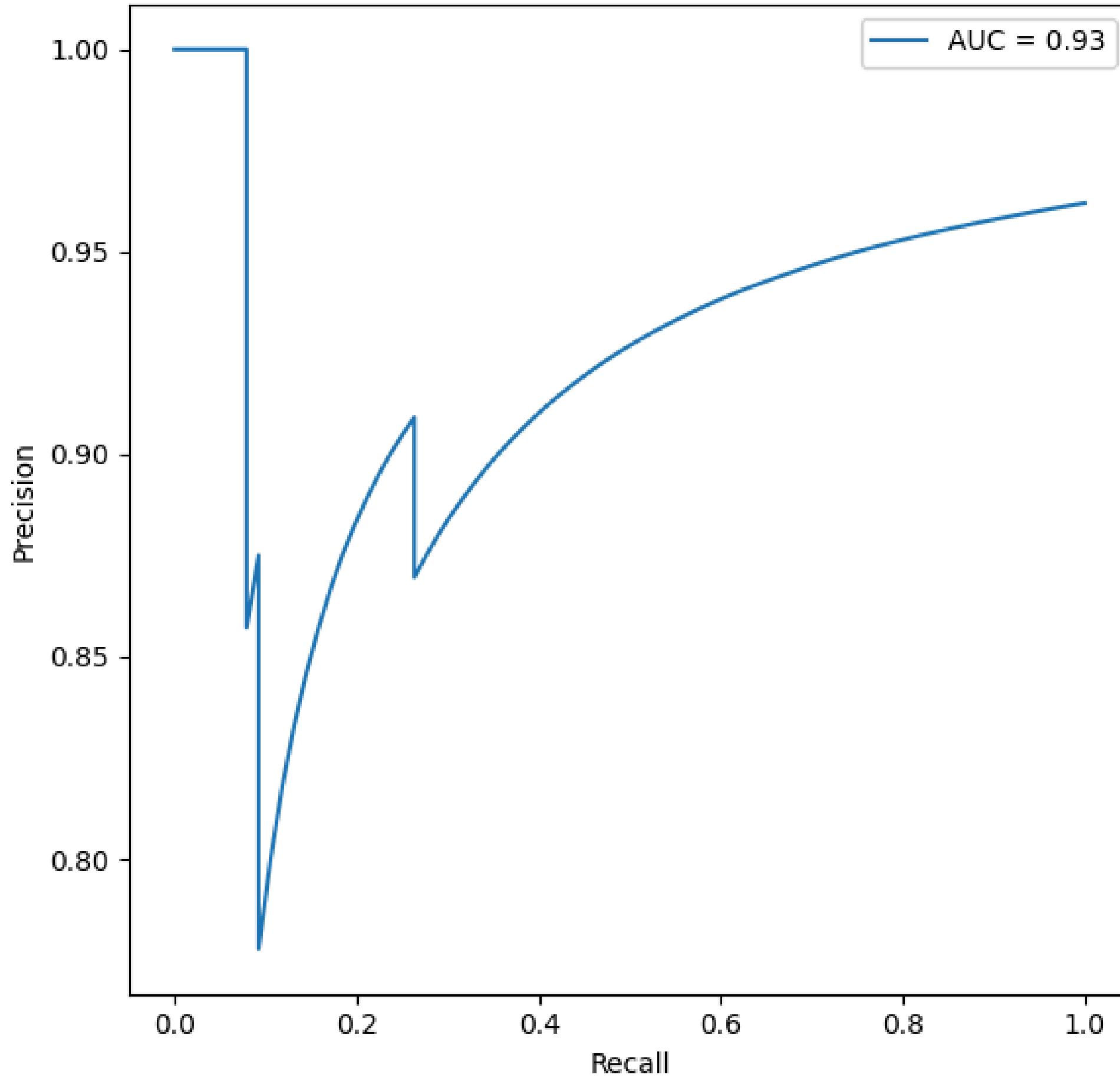
Curva P-R

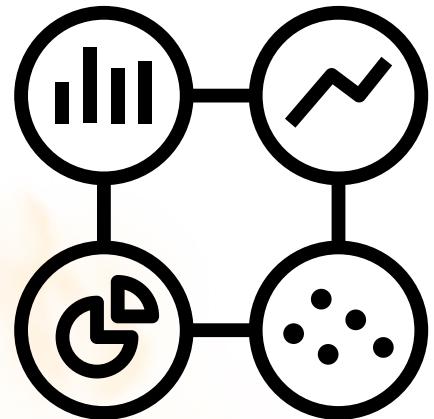


Curva Precision-Recall



Curva Precision-Recall - Modelo Ontológico (Test Set)





Síntesis de la experimentación

- **Modelos tradicionales:** a veces mejor puntuación en métricas específicas.
- **Modelo ontológico:**
 - Atributos organizados de forma coherente y semántica
 - Robusto y equilibrado sin necesidad de fine-tuning para el caso de Predicción de Calificaciones Finales y Clasificación de estudiantes según su desempeño.
 - Mejor capacidad de generalización a nuevos conjuntos de datos en el área educativa: Enriquecer los datos crea **modelos más fiables y transparentes**.





Conclusiones

Integración ontológica en ML

- La ontología enriquece conceptualmente los datos
- Agrupó atributos en categorías como *StudyHabits* o *FamilyContext*
- Aumentó la estructura y explicabilidad del modelo

Consideraciones metodológicas

- Requiere conocimiento del dominio
- Importa el diseño del vocabulario, jerarquías y relaciones
- Es un proceso iterativo

Desempeño en clasificación (Caso 2)

- El modelo ontológico redujo confusiones entre clases
- **Valor agregado:** técnico y pedagógico

Competencias y desafíos

- *Uso de herramientas:* Protégé, OWL, rdflib, owlready2
- Resolución de problemas en integración, razonamiento y validación

Trabajo Futuro

Extensión a otros contextos

- Aplicar en distintas materias, instituciones o países
- Evaluar generalización y adaptar a nuevas tareas (abandono, trayectorias, recomendación)

Evolución de la ontología

- Ampliar con nuevas clases: emociones, digitalidad, extracurriculares
- Usar ontologías existentes y PLN para enriquecer conceptos

Integración de LLMs

- Incorporar LLMs como copilotos semánticos
- Proponer mejoras al mapeo y estructura
- Crear un loop de ajuste dinámico ontología–datos

