UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

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

Daniela Anaí Jiménez Gómez

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniera en Ciencias de la Computación

Quito, 12 de mayo de 2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

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

Daniela Anaí Jiménez Gómez

Nombre del profesor, Título académico **Ricardo Flores, Ph.D**

Quito, 12 de mayo de 2025

© DERECHOS DE AUTOR

Nombres y apellidos: Daniela Anaí Jiménez Gómez

Código: 00322800

Cédula de identidad: 1726862046

Lugar y fecha: Quito, 12 de mayo de 2025

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

El presente trabajo propone una metodología de ingeniería de características basada en ontologías, con el objetivo de mejorar el rendimiento y la interpretabilidad de modelos de machine learning. A través del desarrollo de una ontología educativa construida sobre el dataset "Student Performance", se integraron estructuras semánticas al pipeline de procesamiento, permitiendo una agrupación conceptual y justificada de atributos. Se implementaron y compararon modelos para tres tareas: predicción de calificaciones finales, clasificación del desempeño académico y detección de anomalías, contrastando enfoques tradicionales con versiones enriquecidas por la ontología.

Aunque los modelos estadísticos superaron en métricas cuantitativas, los modelos ontológicos demostraron ser competitivos, estructurados y con mayor trazabilidad. Esta investigación resalta el valor de las ontologías como herramienta complementaria para el aprendizaje automático, especialmente en contextos donde la interpretación y sostenibilidad del conocimiento son fundamentales. El trabajo sienta bases para futuras investigaciones en dominios educativos y otras áreas con alta complejidad semántica.

Palabras clave: ontologías, ingeniería de características, machine learning, representación semántica, rendimiento académico, clasificación, detección de anomalías, educación, análisis de datos.

ABSTRACT

This study proposes an ontology-based feature engineering methodology aimed at enhancing both the performance and interpretability of machine learning models. By developing an educational ontology structured around the "Student Performance" dataset, semantic structures were integrated into the processing pipeline to conceptually group and justify data attributes. Models were implemented and compared across three tasks: final grades prediction, academic performance classification, and anomaly detection contrasting traditional approaches with ontology-enhanced versions.

While statistical models outperformed in quantitative metrics, the ontology-driven models proved competitive, more structured, and offered greater traceability. This research highlights the potential of ontologies as a complementary tool for machine learning, particularly in contexts where interpretability and knowledge sustainability are crucial. The project establishes a foundation for further applications in educational domains and other areas involving high semantic complexity.

Key words: ontologies, feature engineering, machine learning, semantic representation, academic performance, classification, anomaly detection, education, data analysis.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El crecimiento exponencial de datos en las organizaciones exige metodologías avanzadas para abordar su diversidad y complejidad. Por otro lado, la eficacia de los modelos de *machine learning* depende en gran medida de la calidad de las características utilizadas, pero su preparación suele ser manual y propensa a errores. Las técnicas de *feature engineering* y el *data preprocessing* son esenciales para mejorar el rendimiento de los modelos, y en este contexto, el uso de ontologías ofrece un enfoque estructurado para automatizar la selección y generación de características, facilitando una representación semántica más precisa de los datos. Al capturar relaciones conceptuales y significado dentro de un dominio, las ontologías permiten enriquecer los datos de manera más eficiente y reducir la ambigüedad en su interpretación. Este trabajo explora cómo las ontologías pueden optimizar la ingeniería de características y mejorar el desempeño de los modelos de machine learning. Se discutirán los beneficios de este enfoque, sus desafíos y su aplicabilidad en distintos escenarios. A continuación, se presenta el marco teórico que sustenta esta propuesta.

ESTADO DEL ARTE

Ingeniería de características en machine learning

La ingeniería de características es un proceso esencial en el desarrollo de modelos de *machine learning*, encargado de transformar datos crudos en un formato óptimo para su procesamiento por algoritmos predictivos [1]. Mediante la extracción, selección y transformación de características relevantes, se puede mejorar significativamente la precisión y eficiencia de los modelos [2]. Sin embargo, este proceso puede ser complejo, laborioso y propenso a errores si no se maneja de manera estructurada. La calidad de las características influye directamente en el rendimiento de los modelos, desde simples regresiones lineales hasta arquitecturas avanzadas como Redes Neuronales Convolucionales (CNNs) y Redes Neuronales Recurrentes (RNNs) [3]. Mientras que las CNNs destacan en la identificación de patrones espaciales en grandes volúmenes de datos, las RNNs son especialmente eficaces en el procesamiento de secuencias, como el lenguaje natural y las series temporales. Por ello, contar con estrategias sistemáticas para la ingeniería de características resulta fundamental para optimizar el desempeño de estos modelos.

Ontologías como soporte para la ingeniería de características

En este contexto, las ontologías se consolidan como recursos clave para estructurar y representar el conocimiento de manera semántica [3]. Se definen como representaciones formales y explícitas de los conceptos relevantes en un dominio específico, así como de las relaciones existentes entre ellos. Estas estructuras permiten modelar de forma clara entidades, propiedades e interacciones, facilitando una comprensión común entre humanos y sistemas computacionales. Gracias a esta semántica compartida, es posible integrar, estandarizar y enriquecer datos de forma automatizada, habilitando mecanismos de razonamiento que

mejoran la calidad de la información utilizada por los modelos. A continuación, se muestra un ejemplo de una ontología simple empleada en el ámbito de la robótica, tomada de Sun et al. (2019):

Figura 1. Visualización de una ontología usada en el artículo académico *RTPO: A domain knowledge base for robot task planning* [9]

Al aplicar ontologías en la ingeniería de características, los científicos de datos pueden asegurar una mayor coherencia y precisión en la interpretación de los datos, alineando las características extraídas con definiciones y relaciones semánticas claras y bien definidas [4]. Entre las herramientas más utilizadas para gestionar ontologías se encuentran LOV (Linked Open Vocabularies), que proporciona un conjunto central de vocabularios para promover el

uso de datos enlazados; Ontology Lookup Service, que ofrece plataformas diseñadas para facilitar la integración de conocimiento estructurado en procesos de análisis de datos; y BioPortal, que ofrece acceso a una amplia biblioteca de ontologías para el ámbito biomédico. Los modelos de machine learning, desde simples regresiones lineales hasta complejas redes neuronales, se benefician directamente de una ingeniería de características efectiva.

La incorporación de estructuras ontológicas en el proceso de ingeniería de características permite una organización más sistemática de los datos, facilitando su interpretación y reduciendo la ambigüedad semántica. Este enfoque es especialmente útil en contextos donde los datos son heterogéneos o presentan relaciones complejas, ya que las ontologías proporcionan un marco explícito para definir categorías, jerarquías y vínculos entre atributos. Al generar representaciones enriquecidas y coherentes, se mejora la trazabilidad de las variables utilizadas, se reduce la redundancia y se promueve la reutilización del conocimiento. Así, las ontologías no solo optimizan el proceso de preparación de datos, sino que también contribuyen a la explicabilidad de los modelos y a su adaptación a nuevos dominios.

Aplicación de ontologías en el ámbito biomédico

El uso de ontologías en la ingeniería de características ha demostrado ser particularmente valioso en la investigación biomédica, donde la complejidad y heterogeneidad de los datos presentan un gran desafío para el *machine learning*. Un estudio representativo es el de Faust et al. (2019), quienes aplicaron técnicas avanzadas de ingeniería de características combinadas con mapeo ontológico para analizar histomorfologías de tumores cerebrales mediante *deep learning*. Este enfoque mejoró significativamente la precisión en la clasificación de los tipos de tumores y facilitó una interpretación más detallada de los datos biomédicos,

demostrando la capacidad de las ontologías para estructurar la información en modelos predictivos [5].

De manera similar, Kulmanov et al. (2020) exploraron cómo las ontologías, en combinación con técnicas de *machine learning*, pueden impulsar la investigación biomédica. Su estudio destaca el papel de las ontologías en el análisis de grandes volúmenes de datos, ayudando a identificar relaciones biológicas y nuevos biomarcadores de manera más eficiente y precisa [8].

Por otro lado, Sahoo et al. (2022) implementaron ontologías en flujos de trabajo de *machine learning* para gestionar registros clínicos de pacientes con epilepsia. La incorporación de estas estructuras permitió una caracterización más precisa y semánticamente enriquecida de los datos, facilitando diagnósticos y tratamientos personalizados. Su enfoque resalta cómo las ontologías pueden ser utilizadas para manejar la diversidad y complejidad de los datos clínicos, mejorando la personalización del tratamiento y la precisión diagnóstica [7].

Ontologías en el procesamiento de lenguaje natural y análisis de sentimientos

Además del ámbito biomédico, las ontologías han sido aplicadas en el procesamiento de lenguaje natural (NLP) para mejorar la extracción y análisis de características en modelos de *machine learning*. Un ejemplo clave es el trabajo de Siddiqui et al. (2019), quienes desarrollaron un modelo de análisis de sentimientos basado en ontologías. Su investigación demostró cómo las ontologías pueden capturar y expresar características semánticas de los datos de opinión, lo que resultó en mejoras significativas en la precisión del análisis de sentimientos. Al estructurar el conocimiento y definir relaciones conceptuales dentro del lenguaje, el uso de ontologías permitió reducir la ambigüedad en la interpretación de textos y mejorar la clasificación de sentimientos en grandes volúmenes de datos [6].

Impacto y perspectivas del uso de ontologías en machine learning

Estos estudios evidencian el potencial de las ontologías para mejorar la ingeniería de características en distintos dominios de aplicación. Desde el análisis biomédico hasta el procesamiento del lenguaje natural, la capacidad de estructurar datos de manera semántica permite mejorar la calidad de las características extraídas y, en consecuencia, optimizar el rendimiento de los modelos de *machine learning*. A medida que las técnicas de ontología siguen evolucionando, se espera que su adopción en *machine learning* continúe expandiéndose hacia nuevas áreas, facilitando modelos más interpretables, eficientes y precisos.

DESCRIPCIÓN DE LA PROPUESTA

Este trabajo propone integrar ontologías en la ingeniería de características para optimizar este proceso, automatizando la estandarización y enriquecimiento de datos mediante una estructuración semántica. El uso de ontologías permite una clasificación y etiquetado automático que mejora la integridad y utilidad de los datos para el entrenamiento de modelos. Al desarrollar un prototipo experimental que aplica ontologías para la mejora del rendimiento de modelos de ML, el proyecto busca tanto elevar la precisión y eficacia de los modelos como contribuir a la estandarización de prácticas de preparación de datos, resultando en un impacto significativo dentro del ámbito académico y en la aplicación industrial.

Contexto del problema: Modelado inteligente del rendimiento académico estudiantil

En el ámbito educativo, uno de los desafíos más importantes es predecir el rendimiento académico de los estudiantes. Esta información resulta clave para que las instituciones

educativas puedan identificar a tiempo a aquellos alumnos que podrían necesitar apoyo adicional, y así diseñar estrategias de intervención más eficaces.

Para abordar este problema, se utiliza el conjunto de datos ***"Student Performance"***, disponible públicamente a través del UCI Machine Learning Repository. Este dataset reúne información proveniente de dos escuelas portuguesas de nivel secundario y fue recopilado mediante reportes escolares y cuestionarios. Contiene un total de 649 instancias y 33 atributos, que incluyen aspectos académicos, demográficos, familiares, sociales y escolares.

Por tanto, el trabajo se enfocará en el análisis de esta información para modelar distintos tipos de desempeño estudiantil en la asignatura de matemáticas. A fin de enriquecer este análisis, se incorpora una ontología que permita organizar y categorizar los atributos del conjunto de datos de forma estructurada y semánticamente coherente. Esto habilita un enfoque de ingeniería de características más robusto, que facilita la interpretación y puede contribuir a una mejora en la precisión de los modelos predictivos implementados.

Metodología de trabajo

1. *Diseño y desarrollo de ontologías:* Se analiza el conjunto de datos en "Student Performance" y se identifican los principales conceptos y atributos relevantes para representar semánticamente la información contenida. A partir de este paso, se construye una primera versión gráfica de la ontología agrupando las variables en

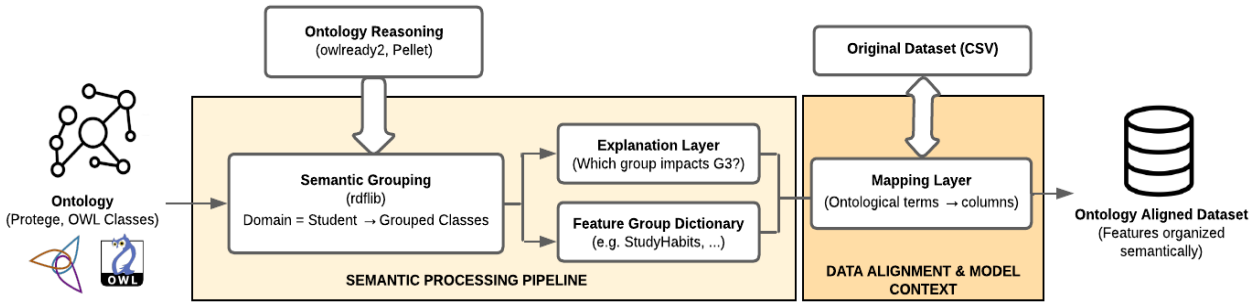
dominios significativos, los cuales facilitan el diseño estructurado en OWL (Web Ontology Language), un lenguaje formal propuesto por el W3C que permite definir clases, relaciones y propiedades con una semántica bien definida, lo que resulta adecuado para el razonamiento automático y la interoperabilidad entre sistemas [10].

2. *Desarrollo del prototipo:* Implementación de la ontología en Protégé [1], una herramienta de código abierto ampliamente utilizada para la creación, edición y visualización de ontologías en OWL, que permite modelar estructuras semánticas complejas de manera intuitiva. Se empleó el plugin OntoGraf para definir jerarquías de clases y relaciones entre entidades clave como Student, Family, AcademicPerformance, entre otras. La ontología fue construida en formato .ttl (formato *Turtle – Terse RDF Triple Language*) con base en los elementos del dataset, estructurando explícitamente el conocimiento relacionado al contexto académico de los estudiantes.

3. *Implementación de la ontología en el pipeline:* Integración de la ontología al flujo de procesamiento de datos de machine learning mediante un enfoque de ingeniería de características guiado. Se desarrolla un prototipo experimental capaz de operar tanto con un modelo tradicional (sin conocimiento semántico) como con uno que incorpora las estructuras ontológicas previamente definidas.

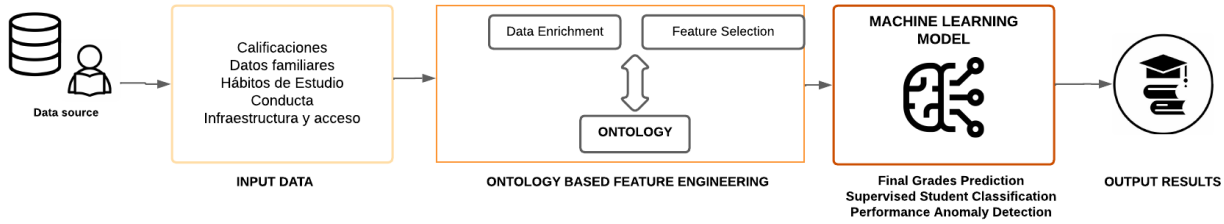
[1] *Protégé.* Open-source ontology editor by Stanford University. Disponible en: https://protege.stanford.edu/

Figura 2. Flujo de procesamiento para enriquecimiento semántico y agrupación explicativa de características mediante ontologías



Nota. Representación del bloque ontológico previo al entrenamiento de un modelo de machine learning, donde se agrupan características según relaciones semánticas definidas en la ontología.

Figura 3. Integración de la ontología en el pipeline de machine learning



Nota. La ontología organiza y enriquece las características antes del entrenamiento, actuando como puente semántico entre los datos y el modelo.

4. *Configuración y pruebas de los experimentos:* Ejecución de distintos experimentos con varios tipos de modelos, incluyendo tareas de predicción, clasificación y detección de anomalías. Se establecen configuraciones paralelas con y sin el uso de la ontología con el fin de evaluar su impacto comparativo sobre el rendimiento de los modelos.

5. *Análisis de resultados:* Análisis comparativo de las métricas obtenidas en los modelos implementados para determinar la utilidad de la ontología en el mejoramiento de los

procesos de inferencia, generalización y comprensión del comportamiento estudiantil desde una perspectiva computacional.

DESARROLLO DEL PROTOTIPO

Diseño inicial de la ontología basada en el dataset *Student Performance*

En esta etapa, el objetivo es diseñar una ontología en el ámbito de la educación orientada a representar de forma estructurada y semántica los conceptos presentes en el conjunto de datos propuesto. Para ello, se realizó un estudio detallado de los atributos disponibles, agrupándolos manualmente en dominios significativos relacionados con la realidad académica y personal del estudiante.

Tabla 1. Agrupación de Features del Dataset en Entidades Ontológicas

Entidad	Features del dataset incluidas
Student	school, sex, age, address (datos generales del estudiante)
Family	famsize, Pstatus, guardian (estructura familiar)
Parents' Occupation	Medu, Fedu, Mjob, Fjob (nivel educativo y trabajo de los padres)
Academic Performance	G1, G2, G3, failures (calificaciones y desempeño académico)
Study Habits	studytime, schoolsup, famsup, paid , nursery, higher (tiempo de estudio y apoyos educativos)

Social Behaviour	goout, activities, romantic, freetime, famrel (interacciones y motivaciones)
Health & Well – being	health, Dalc, Walc, absences (estado de salud y hábitos personales)
Infrastructure	traveltime, internet, reason (condiciones de acceso a la educación)

Entidades Clave

Se identificaron las siguientes entidades clave del dominio educativo, las cuales representan dimensiones fundamentales para el análisis del rendimiento estudiantil. Estas entidades fueron modeladas como clases ontológicas, permitiendo organizar los atributos del dataset de manera semántica y coherente, y facilitando su posterior uso en tareas de machine learning.

1) **Estudiante**: Representa a un estudiante y sus características personales, como edad, género y dirección.
2) **Familia**: Contiene información sobre la estructura familiar y el nivel educativo de los padres.
3) **Rendimiento Académico**: Registra las calificaciones de los estudiantes en diferentes periodos y su historial de fallos.
4) **Hábitos de Estudio**: Incluye información sobre el tiempo de estudio, apoyos educativos adicionales y expectativas de educación superior.
5) **Comportamiento Social**: Representa actividades extracurriculares, relaciones interpersonales y tiempo libre del estudiante.
6) **Salud y Bienestar**: Considera el estado de salud del estudiante, el consumo de alcohol y la cantidad de ausencias.

7) **Infraestructura y Accesibilidad**: Incluye acceso a internet, tiempo de traslado a la escuela y razones de selección de la institución.
8) **Ocupación de los Padres**: Define el empleo de los padres y su posible impacto en el rendimiento del estudiante.

Esquema ontológico

A partir de la identificación de las entidades clave, se diseñó el esquema ontológico que establece las relaciones entre estas entidades por medio de Onto4All[2], la cual es una herramienta para modelar y graficar ontologías de forma visual e intuitiva. Además, cabe recalcar que cada entidad fue asociada a propiedades que describen tanto relaciones objetuales (entre clases) como propiedades de datos.

- **Un Estudiante pertenece a una Familia**, lo que permite evaluar el impacto del entorno familiar en su rendimiento académico.
- **Un Estudiante tiene un Rendimiento Académico**, reflejando sus calificaciones y su historial de fallos.
- **Un Estudiante sigue ciertos Hábitos de Estudio**, que pueden influir en sus resultados académicos.
- **Un Estudiante participa en un Comportamiento Social**, lo que puede afectar su rendimiento y bienestar emocional.
- **Un Estudiante tiene un Estado de Salud y Bienestar**, incluyendo factores como consumo de sustancias y número de ausencias.
- **Un Estudiante accede a Infraestructura y Recursos**, como internet o transporte, que pueden facilitar su proceso de aprendizaje.

[2] *Onto4All.* Free graphical editor for building and exporting ontologies. Disponible en: https://onto4all.com/en

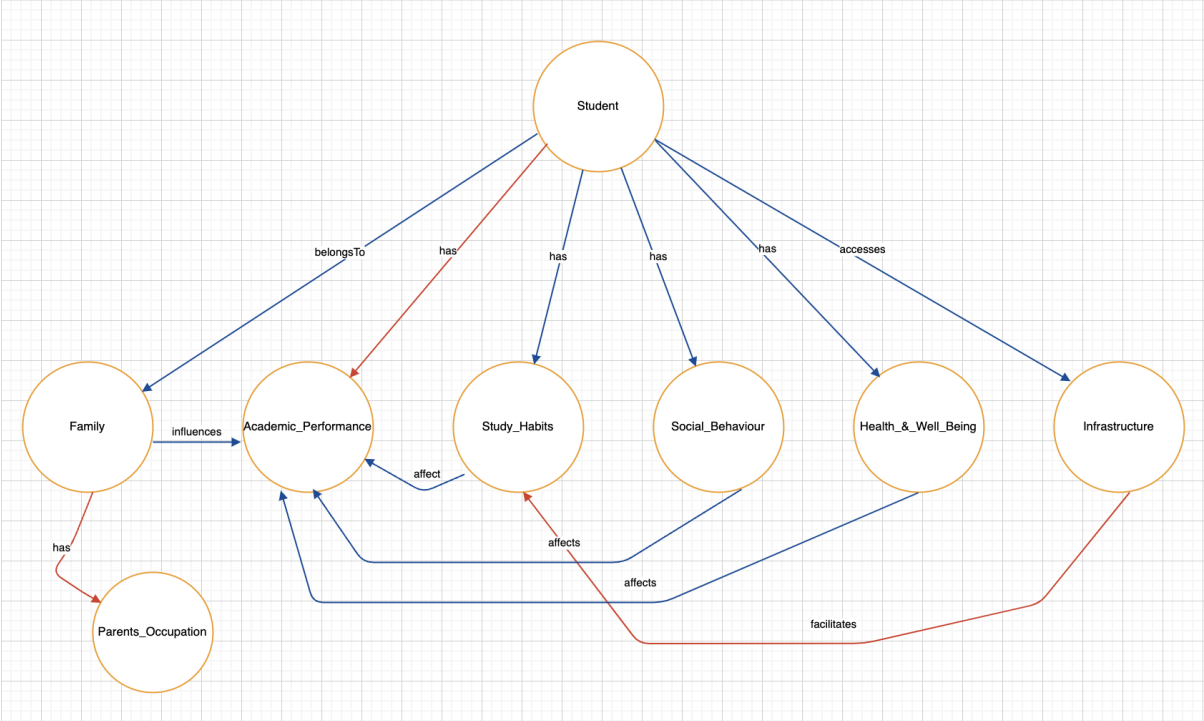
- **Una Familia tiene Ocupación de los Padres**, lo que puede afectar el nivel de apoyo educativo en el hogar.

Tabla 2. Relaciones Definidas en la Ontología del Dataset de Estudiantes

Entidad 1	**Relación**	**Entidad 2**
Student	*belongsTo*	Family
Student	*has*	Academic Performance
Student	*has*	Study Habits
Student	*has*	Social Behaviour
Student	*has*	Health & Well – being
Student	*accesses*	Infrastructure
Family	*has*	Parents' Occupation
Family	*influences*	Academic Performance
Study Habits	*affects*	Academic Performance
Social Behaviour	*affects*	Academic Performance
Health & Well – being	*affects*	Academic Performance
Infrastructure	*facilitates*	Study Habits

Este esquema inicial permitió organizar los datos de manera estructurada, lo cual puede facilitar la interpretación y el enriquecimiento de la información para experimentar con distintos modelos orientados a tareas clave dentro del análisis educativo, como *Predicción de calificaciones finales, Clasificación de estudiantes según su desempeño y Detección de anomalías en el rendimiento educativo*.

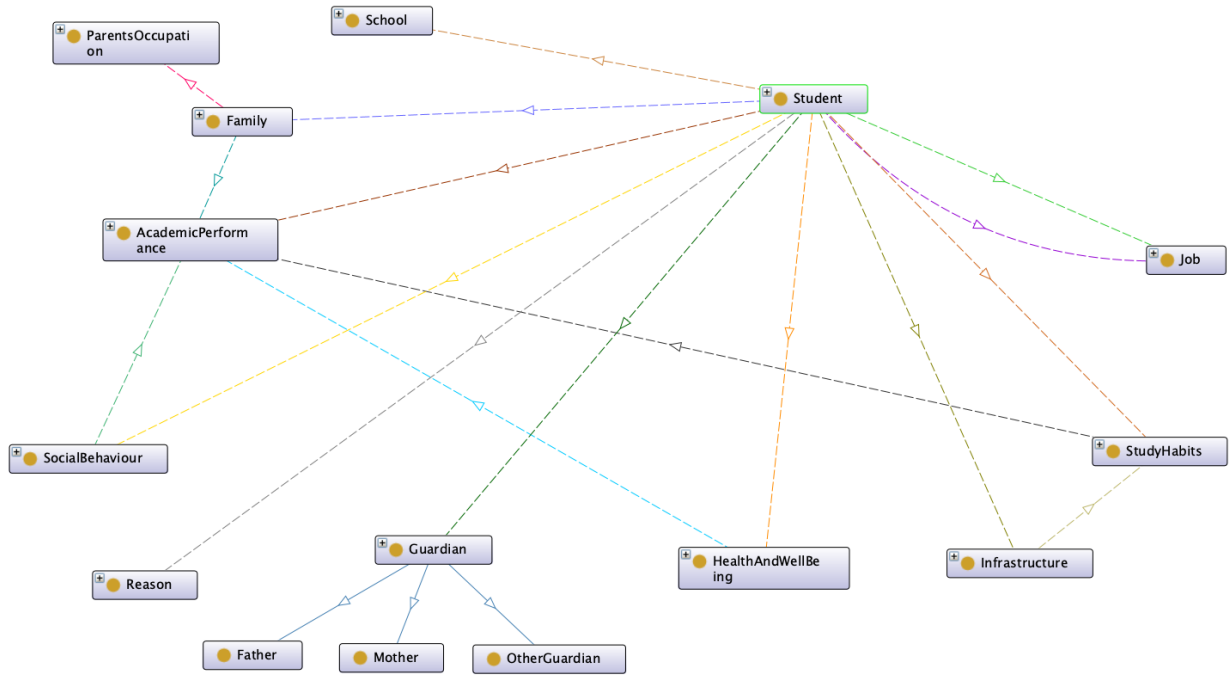
Figura 4. Esquema ontológico inicial del dataset *Student Performance* en Onto4All.



Implementación de la ontología en un pipeline de machine learning

Como parte del desarrollo del prototipo, se redefinió y optimizó el esquema ontológico inicial en Protégé, haciendo uso del plugin *OntoGraf* para visualizar las relaciones entre clases y propiedades. Esto permitió detectar redundancias y ajustar jerarquías conceptuales como la división de guardianes en madre, padre y otro tutor legal. Adicionalmente, se formalizaron conceptos clave como *StudyHabits*, *SocialBehaviour*, *Infrastructure* y *AcademicPerformance*. La ontología quedó representada en un archivo .ttl conforme a OWL, utilizando el formato ttl, el cual permite describir recursos y relaciones en forma de tripletas RDF (Resource Description Framework) de manera legible y compacta, siendo ampliamente utilizado para representar ontologías semánticas en la web.

Figura 5. Ontología refinada para *Student Performance* en Protégé, mostrando relaciones clave entre el estudiante y su contexto.



Entrenamiento experimental de modelos en Python para el análisis educativo.

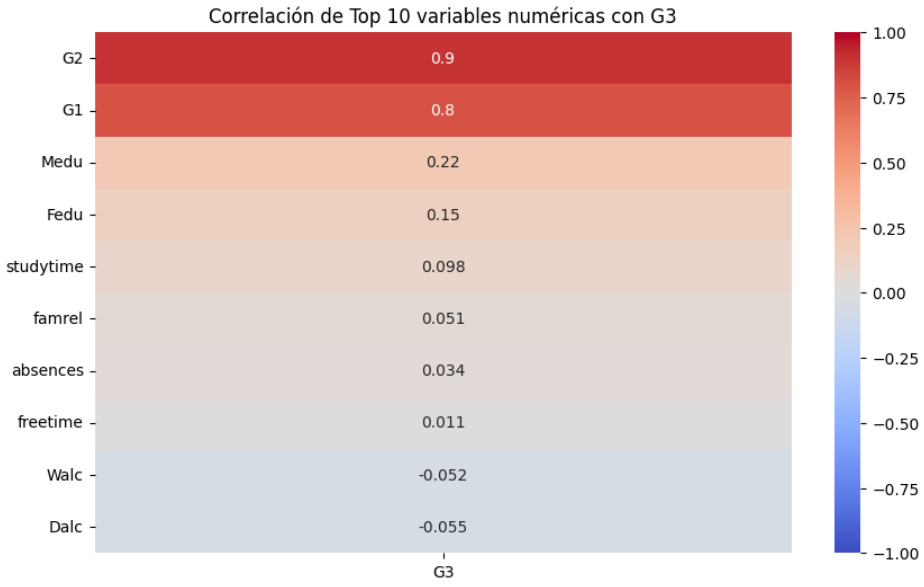
Caso 1: Predicción de calificaciones finales

Con el objetivo de predecir la calificación final de los estudiantes (variable objetivo o target G3), se diseñó un conjunto de experimentos de regresión supervisada utilizando el algoritmo Random Forest Regressor. Este enfoque se centra en predecir un valor continuo (nota final), por lo que se aplicó un proceso de regresión en cada uno de los modelos construidos. Para este caso se compararon dos variantes sin integración ontológica y un modelo ontológico, con el propósito de analizar el impacto de distintas estrategias de selección de características.

Modelos sin integración ontológica

1. **Modelo Base (selección manual de features):** Entrenado con un subconjunto de 8 variables comúnmente reconocidas en la literatura como relevantes para el rendimiento académico. Estas variables fueron seleccionadas por criterio experto e incluyen: calificaciones anteriores (G1, G2), tiempo de estudio (studytime), número de fallos académicos (failures), número de ausencias (absences) y factores socioemocionales como salidas con amigos (goout) y consumo de alcohol (Dalc, Walc). Este conjunto busca representar dimensiones clave sin recurrir a técnicas estadísticas automatizadas.
2. **Modelo Top 10 Features:** Este modelo se construyó tras realizar un análisis de correlación lineal que identificó las 10 variables numéricas con mayor correlación con G3. El objetivo fue comprobar si un proceso de selección automatizado basado en medidas estadísticas puede superar o complementar el rendimiento de un modelo construido con criterio experto.

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



Ambos modelos fueron implementados utilizando la biblioteca *sklearn* de python. El conjunto de datos fue particionado en un 80% para entrenamiento y 20% para prueba, manteniendo la misma semilla aleatoria (*random_state* = 42) para garantizar la reproducibilidad. Los modelos fueron evaluados con métricas estándar de regresión: el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R^2).

Configuración de ambos modelos de regresión:

- *Algoritmo:* RandomForestRegressor
- *Hiperparámetros:* valores por defecto (n_estimators=100, max_depth=None, entre otros)
- *Ajuste de hiperparámetros:* no se realizó, ya que el propósito era mantener una base consistente y controlada para comparar con el modelo ontológico.

A continuación se detallan los resultados obtenidos:

Tabla 3. Comparación de métricas de regresión entre modelos sin integración ontológica para la predicción de calificaciones finales (G3)

Modelo	**MSE**	**R^2**
Modelo Base (8 features)	3.20	0.84
Top 10 features correlacionadas	2.92	0.86

Al analizar las métricas, se identificó que el modelo de regresión con selección automatizada de variables mediante análisis de correlación presentó una mejora leve respecto al modelo construido con criterio experto. Esta mejora se reflejó en valores más bajos de MSE

y una ligera ganancia en el coeficiente R^2. Sin embargo, en un entorno como el educativo, donde la comprensión de los factores que afectan el rendimiento académico es tan importante como la predicción en sí, este incremento técnico no necesariamente justifica el abandono de enfoques guiados por conocimiento del dominio. La capacidad de interpretar y justificar las variables utilizadas continúa siendo una prioridad. Por ello, ambos modelos establecen un marco de referencia útil para evaluar la pertinencia de incorporar enfoques semánticos más estructurados, como el uso de la ontología para organizar las variables bajo categorías conceptuales coherentes.

Implementación de la ontología en el pipeline.

Como se ha mencionado anteriormente, en esta etapa de experimentación el propósito es evaluar el impacto de una representación semántica del dominio educativo en la tarea de predicción de calificaciones finales, por lo cual se integró la ontología previamente desarrollada al flujo de procesamiento del modelo de regresión. Este proceso consistió en aplicar un enfoque de *feature engineering* guiado por conocimiento semántico, reemplazando métodos tradicionales como la selección manual o estadística de features por conocimiento formalizado y variables organizadas según clases ontológicas.

1. **Semantic Feature Grouping: agrupamiento semántico como técnica de ingeniería de características**

En este caso se empleó la técnica de *semantic feature grouping*, que consiste en organizar variables del dataset según las clases y propiedades de una ontología. A diferencia de la feature selection estadística (que elimina variables con poca correlación o importancia), este enfoque no reduce el número de atributos, sino que estructura su inclusión con base en conocimiento semántico explícito.

Para aplicar esta técnica se utilizaron dos herramientas complementarias:

- **rdflib:** Es una librería que permite manipular grafos RDF y consultar relaciones semánticas entre entidades. Para este caso, fue usada para cargar la ontología desarrollada en formato Turtle (.ttl) y se construyó un grafo RDF formado por tripletas de *sujeto, predicado, objeto* que representan las relaciones semánticas. Desde ahí, se extrajeron todas las propiedades cuyo dominio era la clase Student. Estas propiedades se agruparon por su clase de rango, como StudyHabits, HealthAndWellBeing, Family, etc.
- **owlready2:** Es la librería para trabajar con ontologías OWL, que permite cargarlas, editarlas y realizar razonamiento lógico. De forma puntual, esta herramienta se usó para convertir la ontología al formato OWL (.owl) y ejecutar el razonamiento lógico mediante el reasoner Pellet. Esto permitió verificar la consistencia del modelo ontológico, inferir clases y relaciones adicionales, y confirmar que la estructura conceptual usada era válida y semánticamente coherente.

Tras este procesamiento, las propiedades extraídas fueron mapeadas al dataset real, resultando en los siguientes subconjuntos:

1) **AcademicPerformance:** G1, G2, failures
2) **StudyHabits:** studytime, schoolsup, famsup, paid
3) **HealthAndWellBeing:** health, absences, Dalc, Walc
4) **SocialBehaviour:** goout, romantic, freetime
5) **Infrastructure:** internet, traveltime
6) **Family:** famsize, Pstatus
7) **ParentsOccupation:** Medu, Fedu, Mjob, Fjob

8) **Demographics:** age, sex, address, guardian

Esto permitió construir un conjunto de entrenamiento basado en 26 variables originales agrupadas semánticamente.

2. Entrenamiento del modelo ontológico

Las variables categóricas fueron transformadas mediante one-hot encoding (drop_first=True). Posteriormente, se entrenó un modelo de regresión con RandomForestRegressor usando los mismos parámetros experimentales de los modelos sin ontología que se implementaron antes:

Configuración del modelo:

- *Algoritmo:* RandomForestRegressor
- *División:* 80% entrenamiento y 20% prueba
- *Semilla:* random_state=42
- *Hiperparámetros:* valores por defecto (n_estimators=100, max_depth=None, entre otros)
- *Ajuste de hiperparámetros:* no se realiza para tener una comparación justa entre modelos, dado que todos parten de la misma base.

Los resultados obtenidos de la integración ontológica fueron los siguientes:

Tabla 4. Métricas obtenidas del modelo ontológico para predicción de calificaciones finales (G3)

Modelo	**MSE**	**R^2**
Modelo Ontológico	3.54	0.83

Comparación final de modelos y estrategias para la predicción de calificaciones finales

Luego de implementar las distintas alternativas de modelado, se presenta un resumen de los tres enfoques implementados para la predicción de calificaciones finales, destacando sus respectivas técnicas de ingeniería de características, el número de variables utilizadas y su desempeño en términos de métricas:

Tabla 5. Comparación de rendimiento entre modelos predictivos con y sin integración ontológica

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

Aunque el modelo basado en las 10 variables más correlacionadas con la nota final obtuvo el mejor desempeño en términos de MSE y R^2, la diferencia frente al modelo manual fue moderada, lo que refuerza la validez del conocimiento experto para representar factores críticos en el rendimiento académico.

Por su parte, el modelo ontológico presentó métricas ligeramente menores, pero introdujo una perspectiva estructural distinta al organizar las variables mediante agrupamiento semántico. Esta integración semántica no solo permitió incorporar el conocimiento del dominio en el proceso de modelado, sino que funcionó como una forma de estructuración explícita de las características usadas para entrenar los modelos.

A pesar de no haber realizado un ajuste fino de hiperparámetros, el modelo mantuvo un rendimiento competitivo y estable, lo cual sugiere que una representación formal del contexto educativo puede operar como una forma de regularización interpretativa que facilita el modelado en entornos donde la transparencia y la alineación con criterios pedagógicos son igual de importantes que la precisión numérica.

Caso 2: Clasificación de estudiantes según su desempeño académico

Para explorar el rendimiento de otros modelos dentro del mismo contexto educativo, se decidió clasificar a los estudiantes en función de su rendimiento académico. Para ello, se desarrollaron modelos de clasificación supervisada utilizando el algoritmo RandomForestClassifier.

La variable objetivo fue construida a partir de la calificación final (G3), que en el sistema educativo de Portugal (país del cual proviene el dataset utilizado "*Student Performance*"), se evalúa en una escala numérica de 0 a 20 puntos. De acuerdo con el marco normativo del sistema de enseñanza secundaria portugués, una calificación igual o inferior a 9.5 es considerada insuficiente o "reprovado", mientras que a partir de 10 se considera aprobado. Basado en esta escala y respaldado por estudios previos que segmentan el

rendimiento estudiantil en tres niveles según Cortez & Silva (2008)., se definieron los siguientes rangos [11] :

- *Bajo* ($G3 \leq 10$)
- *Medio* ($11 \leq G3 \leq 14$)
- *Alto* ($G3 \geq 15$).

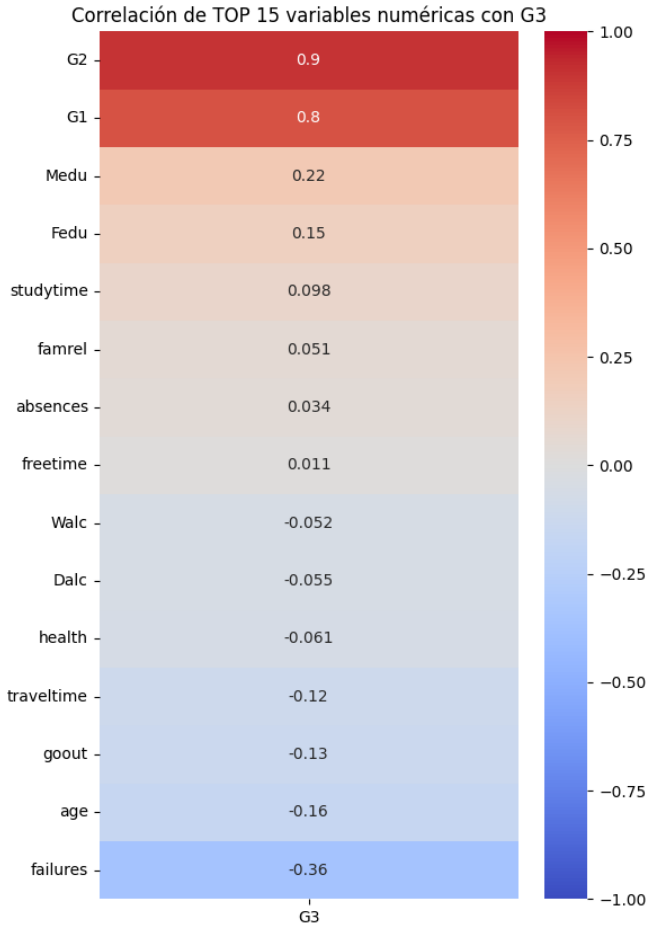
Este planteamiento permitió transformar el problema de regresión previa en una tarea de clasificación multiclase. De forma similar al caso anterior sobre predicción de calificaciones, se implementaron y compararon dos modelos sin integración ontológica y un modelo con integración semántica basada en la misma ontología.

Modelos sin integración ontológica

1. **Modelo Base (selección manual de features):** Este modelo fue entrenado utilizando un subconjunto de 8 variables comúnmente reconocidas en la literatura por su relación con el rendimiento académico. Las variables fueron seleccionadas por criterio experto e incluyeron dimensiones cognitivas y socioemocionales: calificaciones anteriores (G1, G2), tiempo dedicado al estudio (studytime), historial de fracasos (failures), número de ausencias (absences), y factores conductuales como salidas sociales (goout) y consumo de alcohol en días de semana y fines de semana (Dalc, Walc). Este enfoque busca representar un perfil estudiantil equilibrado sin depender de técnicas estadísticas automáticas.
2. **Modelo Top 15 Features:** Este modelo fue construido tras realizar un análisis de correlación lineal con respecto a la variable continua G3. A partir de este análisis se seleccionaron las 15 variables numéricas más correlacionadas, tanto positiva como negativamente, con el rendimiento académico. Si bien el modelo final utiliza una

variable categórica de desempeño como objetivo (*Desempeño*), este paso previo permitió explorar la utilidad de la correlación con G3 como guía para la selección de características. El objetivo fue comprobar si una selección automatizada basada en medidas estadísticas podía superar o complementar la capacidad predictiva del modelo manual.

Figura 7. 15 variables numéricas más correlacionadas con la calificación final (G3).



Es importante mencionar que ambos modelos fueron diseñados manteniendo una configuración consistente, lo cual permitió una comparación justa frente al modelo ontológico posterior. La partición para el conjunto de datos fue de 75% para entrenamiento y 25% para prueba, mientras la semilla aleatoria es (*random_state* = 42). Además, se evaluó el desempeño

de los modelos con las métricas de Precision, Recall, F1-Score y Accuracy, donde este último valor representa el porcentaje de estudiantes correctamente clasificados en su categoría real de desempeño (alto, medio o bajo).

Configuración de ambos modelos de clasificación:

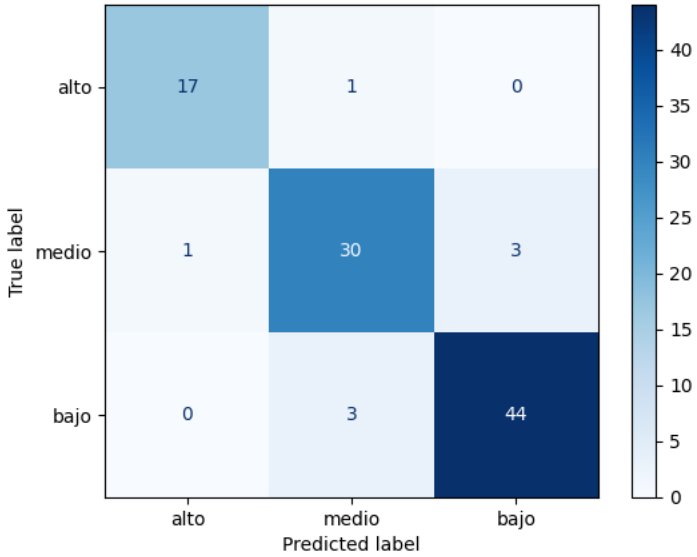
- *Algoritmo:* RandomForestClassifier
- *Hiperparámetros:* valores por defecto (n_estimators=100, criterion='gini', max_depth=None, min_samples_split=2, entre otros)
- *Ajuste de hiperparámetros:* no se aplicó ajuste fino para mantener un punto de referencia en la comparación con el modelo de integración ontológica.

En concordancia con lo anterior, se presentan las métricas de evaluación obtenidas para ambos modelos:

Tabla 6. Métricas obtenidas del modelo de clasificación de desempeño estudiantil con selección manual de features

Categoría	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Alto	0.94	0.94	0.94	0.92
Medio	0.88	0.88	0.88	
Bajo	0.94	0.94	0.94	

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

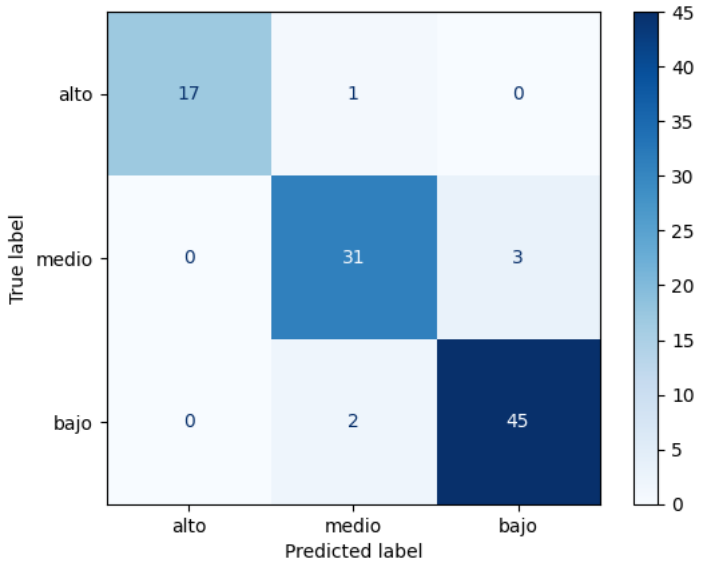


Nota. Se observa una confusión moderada entre las clases “medio” y “bajo”, lo cual indica que estas categorías comparten patrones similares cuando se utilizan variables seleccionadas por criterio experto.

Tabla 7. Métricas obtenidas del modelo de clasificación de desempeño estudiantil con features más correlacionadas

Categoría	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Alto	1.00	0.94	0.97	0.94
Medio	0.91	0.91	0.91	
Bajo	0.94	0.96	0.95	

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.

Como se observa, el modelo con características seleccionadas automáticamente por su mayor correlación con la calificación final G3 obtuvo un rendimiento ligeramente superior en todas las métricas clave frente al modelo con selección manual. Sin embargo, esta diferencia fue marginal, lo cual resulta particularmente relevante considerando el contexto educativo.

La clasificación de estudiantes según su nivel de desempeño (alto, medio o bajo) es una tarea sensible, donde no solo interesa la precisión del modelo, sino también la comprensibilidad y justificabilidad de las variables utilizadas. En ese sentido, el modelo basado en criterio experto sigue siendo válido, especialmente cuando se cuenta con conocimiento del dominio que permita fundamentar la selección de variables.

Ambos modelos, por tanto, ofrecen una base sólida para contrastar con la propuesta ontológica que se presenta a continuación, la cual busca mantener un desempeño competitivo

mientras incorpora beneficios complementarios como trazabilidad semántica, interpretabilidad estructurada y potencial de generalización a otros contextos educativos.

Entrenamiento del modelo ontológico

Para complementar los enfoques tradicionales de clasificación, se implementó un modelo guiado por integración semántica basada en la ontología, replicando la misma lógica de diseño utilizada en el primer caso de estudio (Predicción de calificaciones finales mediante regresión). Esto permitió explorar si la estructuración del conocimiento mediante una ontología puede mantener, o incluso mejorar, el rendimiento de un modelo que ahora realiza una tarea de clasificación.

Proceso de integración semántica.

Dado que se replicó la metodología de integración ontológica utilizada en el caso 1, los pasos para este proceso fueron los siguientes:

- **Carga de la ontología**: Se utilizó el archivo .ttl de la ontología, cargado mediante rdflib y luego convertido a formato OWL para su procesamiento con Owlready2.
- **Razonamiento automático**: Se empleó el reasoner Pellet para inferir propiedades semánticas y validar relaciones entre clases, permitiendo obtener una visión enriquecida del dominio educativo.
- **Semantic Feature Grouping**: A partir de las propiedades del dominio Student, se identificaron grupos conceptuales como hasPerformance, hasStudyHabits, hasHealthProfile y hasSocialBehaviour. Estas propiedades fueron mapeadas a columnas reales del dataset original, resultando en un total de *16 variables* ontológicamente agrupadas.

- **Reducción automática de características**: Como paso adicional, se utilizó SelectFromModel con RandomForestClassifier para realizar una selección automática de las características más importantes dentro del conjunto semántico previamente estructurado.

Configuración del modelo:

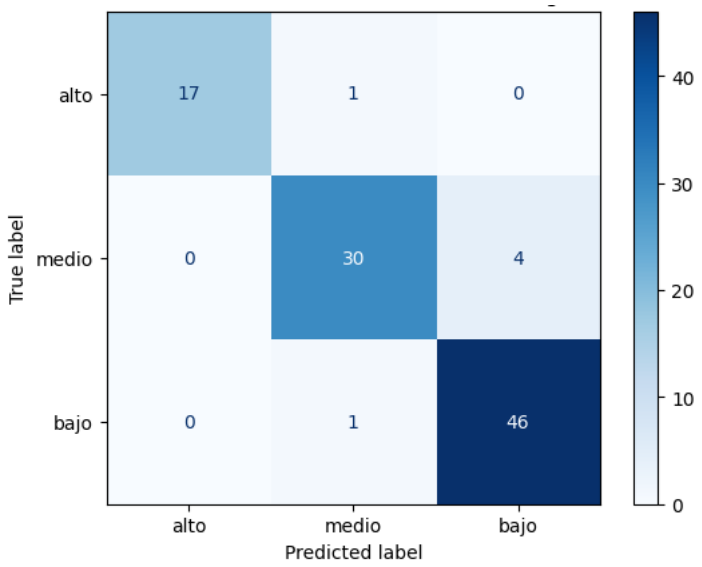
- *Algoritmo:* RandomForestClassifier
- *División:* 75% entrenamiento y 25% prueba (estratificada)
- *Semilla:* random_state=42
- *Hiperparámetros:* valores por defecto (n_estimators=100, criterion='gini', max_depth=None, min_samples_split=2, entre otros)
- *Ajuste de hiperparámetros:* No se realizó.

Los resultados en métricas fueron:

Tabla 8. Métricas obtenidas del modelo ontológico para la clasificación del desempeño estudiantil

Categoría	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Alto	1.00	0.94	0.97	0.939
Medio	0.94	0.88	0.91	
Bajo	0.92	0.98	0.95	

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.

Comparación final de modelos y estrategias para la clasificación de estudiantes según su desempeño académico

Se presenta una síntesis de los tres enfoques implementados para clasificar a los estudiantes en niveles de desempeño. En esta comparación, se optó por destacar particularmente el *F1-score* por clase, ya que esta métrica representa el equilibrio entre precisión y recall, resultando especialmente útil en contextos de clasificación multiclase como este en donde las categorías están desbalanceadas. Asimismo, se incluye la métrica de accuracy general para una visión agregada del rendimiento:

Tabla 9. Comparación de rendimiento entre modelos predictivos con y sin integración ontológica

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

En esta segunda tarea, el modelo con las 15 variables más correlacionadas alcanzó la mayor precisión global, mientras que el modelo ontológico logró un rendimiento muy cercano y altamente equilibrado entre las clases, superando incluso al modelo estadístico en el F1-score de la clase “bajo” y empatando en la clase “alto”. Este comportamiento balanceado resulta especialmente relevante en contextos educativos donde la equidad entre perfiles estudiantiles es crucial.

A diferencia de los modelos base, que se sustentan en heurísticas estadísticas o selección manual, el modelo ontológico empleó una estructura semántica para organizar y reducir dimensionalmente el conjunto de características. Esta representación guiada no solo facilitó la identificación de variables relevantes, sino que también propició una configuración más robusta sin requerir optimización adicional. Al evitar un sobreajuste al dataset y favorecer una segmentación clara por dimensiones del dominio, el modelo ontológico demostró que una ingeniería de características formalizada puede ser tan efectiva como las técnicas tradicionales, especialmente en tareas de clasificación con categorías intermedias y balance desigual.

Caso 3: Detección de anomalías en el rendimiento educativo

Este tercer caso de estudio abordó una problemática distinta a las anteriores: la identificación de estudiantes con bajo rendimiento académico utilizando enfoques de detección de anomalías. A diferencia de los casos previos, esta tarea fue tratada como un problema no supervisado, al no existir una etiqueta directa en el conjunto de datos original para este tipo de casos. Por tanto, se generó una nueva variable objetivo (target), categorizando como anomalía (valor -1) a aquellos estudiantes que habían reprobado más de dos asignaturas (failures > 2), mientras que el resto fue considerado normal (valor 1).

Nuevamente, se exploraron y compararon los tres enfoques estándar para este trabajo: dos modelos sin integración ontológica y un modelo guiado por ontología.

Modelos sin integración ontológica

1. **Modelo Base (selección manual de features + ajuste fino):** Este modelo fue entrenado con 8 variables más asociadas con la posibilidad de reprobar seleccionadas por criterio expert y reconocidas por su relevancia para describir el desempeño académico y el contexto social del estudiante. Estas variables incluyen: G1, G2, studytime, failures, absences, goout, Dalc y Walc. Se estandarizaron las variables mediante StandardScaler y se utilizó el algoritmo Isolation Forest con optimización de hiperparámetros a través de GridSearchCV, empleando un scorer personalizado basado en el recall de la clase anómala. El modelo fue entrenado y validado en el mismo conjunto completo (sin separación explícita entre entrenamiento y prueba) para maximizar la detección de casos minoritarios.

2. **Modelo Top 10 Features (selección estadística + fine tuning):** Este modelo partió de un análisis de correlación entre las variables numéricas y la variable objetivo binaria target, seleccionando automáticamente las 10 características más correlacionadas (positiva o negativamente). Luego, se dividió el conjunto de datos en 80% entrenamiento y 20% prueba (estratificado). Se aplicó escalamiento de variables y optimización de hiperparámetros del algoritmo en el conjunto de entrenamiento, utilizando nuevamente un scorer basado en recall para la clase anómala.

Figura 11. 10 variables numéricas más correlacionadas con 'target'.

Es necesario subrayar que, al igual que en el caso 2 de clasificación, los modelos fueron evaluados mediante las métricas de precisión, recall, F1-score y accuracy, siendo esta última el porcentaje de estudiantes correctamente clasificados como -1 (reprobados) o 1 (no reprobados). También se reportaron dos métricas de área bajo la curva: el AUC-ROC, calculado directamente desde el clasificador para medir la capacidad general de discriminación,

y el AUC Precision-Recall, derivado de la curva P-R, más representativo en escenarios con clases desbalanceadas como el presente caso de detección de anomalías.

Configuración común de ambos modelos:

- *Algoritmo:* IsolationForest
- *Optimización:* GridSearchCV con recall_anomaly (recall negativo o scorer que penaliza falsos negativos)
- *Hiperparámetros ajustados:*
    a. Modelo con selección manual*:* n_estimators=200, max_samples='auto', contamination=0.05
    b. Modelo con top features: n_estimators=100, max_samples='auto', contamination=0.1
- *Ajuste de hiperparámetros:* Sí, aplicado en ambos modelos
- *Semilla:* random_state = 42
- *Escalado:* StandardScaler

Tras las implementación, se obtuvieron los resultados siguientes:

Tabla 10. Métricas obtenidas del modelo de detección de anomalías en el rendimiento educativo con selección manual de features

Clase	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**AUC ROC**	**AUC PR**
-1 (anómalo)	0.25	0.94	0.39	0.88	0.91	0.88
1 (normal)	1.00	0.88	0.94			

Figura 12. Curva Precision-Recall del modelo base (selección manual de variables) para detección de anomalías.

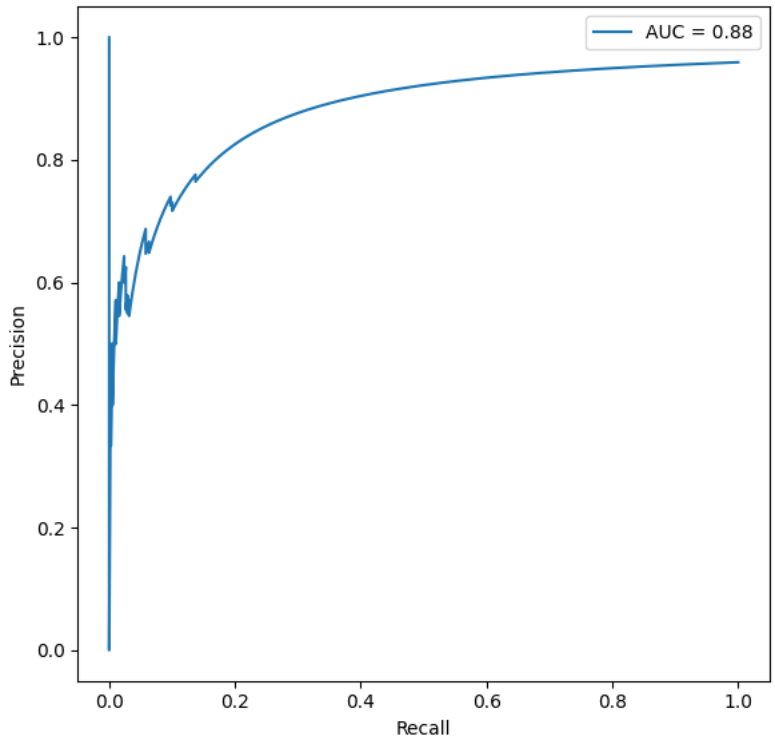
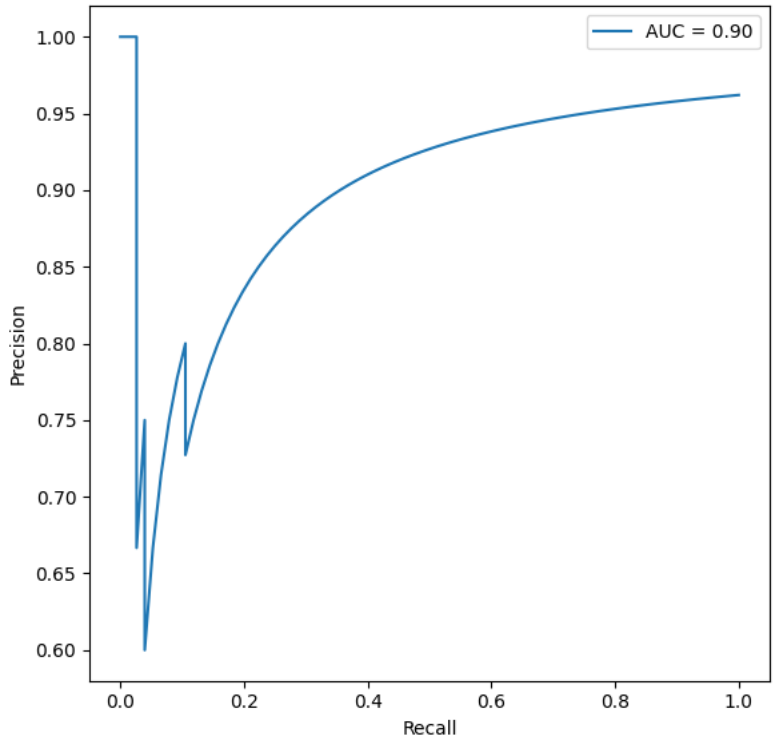


Tabla 11. Métricas obtenidas del modelo de detección de anomalías en el rendimiento educativo con features más correlacionadas

Clase	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**AUC ROC**	**AUC PR**
-1 (anómalo)	0.22	0.67	0.33	0.90	0.79	0.90
1 (normal)	0.99	0.91	0.95			

Figura 13. Curva Precision-Recall del modelo con top 10 variables más correlacionadas para detección de anomalías.



Pese a que fueron optimizados mediante técnicas como Grid Search, los modelos tradicionales presentan rendimientos dispares frente a la detección de estudiantes con bajo rendimiento. El modelo basado en selección manual mostró una alta sensibilidad (recall) para la clase anómala, aunque sacrificando precisión, lo que indica que detectó la mayoría de los casos relevantes, pero con varios falsos positivos.

Por otro lado, el modelo construido con las variables más correlacionadas obtuvo una mayor precisión general y una mejor área bajo la curva Precision-Recall (AUC P-R), lo que sugiere una mejor capacidad de discriminación en general, aunque detectó menos anomalías (recall para -1).

Esta comparación resalta el dilema entre sensibilidad y precision, donde la identificación correcta de estudiantes en riesgo es crítica. Ambos enfoques, sin embargo, permiten establecer una línea base sólida para contrastar el desempeño del modelo ontológico.

Modelo con integración ontológica

Como primer paso, se replicó la metodología de integración semántica utilizada previamente en el Caso 1 (Predicción de calificaciones finales) y Caso 2 (Clasificación de estudiantes). El proceso consistió en:

- **Carga de la ontología**
- **Razonamiento automático**
- **Semantic Feature Grouping:** Para esta tarea, se identificaron grupos semánticos clave (e.g., hasPerformance, hasStudyHabits, hasHealthProfile, hasSocialBehaviour) y se mapearon a 16 columnas del dataset.
- **Preprocesamiento:** Se aplicaron transformaciones categóricas y escalado de variables.
- **Entrenamiento del modelo**

Configuración del modelo ontológico:

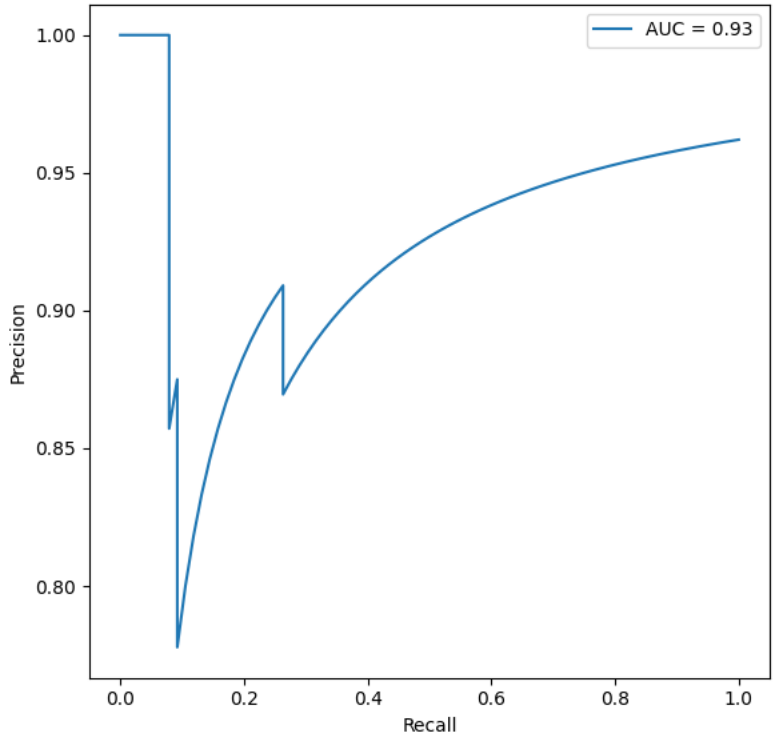
- *Algoritmo:* IsolationForest
- *División:* 80% entrenamiento y 20% prueba (estratificada)
- *Semilla:* random_state = 42
- *Hiperparámetros:* valores por defecto (n_estimators=100, max_samples='auto', contamination=0.1, bootstrap=False)
- *Ajuste fino:* no realizado

Los resultados en métricas fueron:

Tabla 12. Métricas obtenidas del modelo ontológico para detección de anomalías en el rendimiento educativo

Clase	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**AUC ROC**	**AUC PR**
-1 (anómalo)	0.14	0.33	0.20	0.90	0.63	0.93
1 (normal)	0.97	0.92	0.95			

Figura 14. Curva Precision-Recall del modelo ontológico para detección de anomalías.



Comparación final de modelos y estrategias para detección de anomalías

A continuación, se sintetizan los resultados obtenidos por los tres modelos en la tarea de detección de estudiantes con bajo rendimiento académico, destacando las métricas clave mencionadas previamente particularmente para la clase minoritaria (-1), ya que es especialmente útil para el desbalance que se presenta:

Tabla 13. Comparación de rendimiento entre modelos predictivos con y sin integración ontológica

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

Si bien los modelos tradicionales optimizados mediante técnicas estadísticas ofrecieron mejores resultados en métricas como el F1-score o el recall para la clase minoritaria (estudiantes con bajo rendimiento), el modelo ontológico permitió observar un comportamiento singular desde el punto de vista del umbral de decisión. En particular, al no haber sido afinado con hiperparámetros ni validado por técnicas como Grid Search, este modelo mostró una mayor estabilidad frente a variaciones leves en la distribución de anomalías, reflejándose en un valor de AUC-PR notablemente alto.

Esto sugiere que, aunque el modelo no fue el mejor en recall puro, su estructura general logró capturar señales útiles para distinguir patrones atípicos, lo cual es especialmente valioso en escenarios donde el conjunto de anomalías es reducido o se desconoce su proporción real. Además, la configuración fija sin ajuste fino evitó el sobreajuste que puede darse con modelos altamente optimizados sobre datasets pequeños o desbalanceados, mostrando que una representación semántica coherente puede actuar como una forma de regularización implícita.

CONCLUSIONES

Valor del enriquecimiento semántico

Uno de los principales aportes de este estudio radica en haber demostrado que es posible integrar estructuras ontológicas en el flujo de procesamiento de datos para enriquecer conceptualmente las características utilizadas por los modelos. En lugar de trabajar únicamente con atributos planos o seleccionados estadísticamente, la ontología permitió agrupar variables relacionadas en dimensiones más significativas que no se toman en cuenta en primera instancia, como *StudyHabits* o *FamilyContext* lo que aportó mayor explicabilidad y estructura al conjunto de datos.

Desempeño del modelo ontológico

Entre los hallazgos más relevantes se destaca el desempeño del modelo ontológico en la tarea de clasificación de estudiantes según su rendimiento (Caso 2). Este modelo, construido a partir de un agrupamiento semántico de características derivadas de la ontología, logró reducir la confusión entre clases limítrofes, como “medio” y “bajo”, y presentó un rendimiento equilibrado frente a los modelos tradicionales. A pesar de que el enfoque basado en correlaciones estadísticas obtuvo valores ligeramente superiores en métricas como precisión o *recall* en algunos casos, el modelo ontológico destacó por ofrecer mayor coherencia interpretativa, trazabilidad y alineación con el conocimiento experto del dominio educativo. Esto refuerza el valor de las ontologías no solo como herramienta técnica, sino también como vehículo para integrar criterios pedagógicos y semánticos en el aprendizaje automático.

Consideraciones metodológicas

A nivel metodológico, este trabajo evidenció que la implementación de una ontología orientada al dominio requiere una comprensión profunda del contexto y una construcción cuidadosa del vocabulario, las clases y las relaciones. Se identificó que la efectividad del enfoque ontológico depende en gran medida del diseño inicial de la ontología, así como de su capacidad para representar con fidelidad las interacciones y dependencias entre variables del mundo real. Por tanto, el proceso implicó múltiples iteraciones, ajustes y validaciones, reflejando que el desarrollo de este tipo de soluciones es altamente dependiente de la calidad de los datos, del conocimiento experto en el dominio y de una fase experimental extensa que involucra prueba y error.

Competencias desarrolladas y desafíos técnicos

A lo largo del proyecto, se adquirieron valiosas competencias en herramientas como Protégé, OWL, rdflib y owlready2, así como en técnicas de integración entre ontologías y pipelines de aprendizaje automático. También se abordaron desafíos técnicos como el mapeo de términos ontológicos a columnas del dataset, la necesidad de mantener consistencia lógica durante el razonamiento y la validación de resultados en entornos híbridos simbólico-conexionistas.

Recomendaciones y Trabajo Futuro.

Extensión del enfoque a otros contextos educativos

Una línea clara de trabajo futuro consiste en aplicar esta metodología a otros conjuntos de datos relacionados con educación, ya sea en otras materias, instituciones o contextos geográficos. Esto permitiría evaluar la generalización del enfoque y adaptar la ontología a diferentes necesidades educativas, como la predicción de abandono escolar, recomendación de recursos, análisis de trayectorias o segmentación de perfiles estudiantiles.

Evolución de la ontología

La ontología actual puede considerarse una versión base que puede ser ampliada con nuevas clases, propiedades y relaciones. Su evolución podría incluir dimensiones más complejas como factores socioemocionales, hábitos extracurriculares, o interacciones con el entorno digital. También podría beneficiarse del uso de ontologías de dominio existentes o herramientas de enriquecimiento automático basadas en procesamiento de lenguaje natural.

Incorporación de un LLM como experto artificial en el pipeline ontológico

Una línea prometedora de evolución consiste en integrar Modelos de Lenguaje de gran escala (LLMs) como copilotos semánticos dentro del pipeline de machine learning. En particular, se propone construir un sistema de retroalimentación continua basado en técnicas de Recuperación Aumentada por Generación (RAG), donde el LLM actúe como un “experto artificial” capaz de analizar los datos, evaluar la estructura ontológica actual y sugerir mejoras conceptuales. Este loop dinámico permitiría ajustar y refinar la ontología en función del dominio específico del problema, potenciando la adaptabilidad y precisión del modelo. Tal

enfoque facilitaría la evolución automática del conocimiento semántico y su alineación con contextos cambiantes o altamente especializados.

REFERENCIAS

[1] Verdonck, T., Baesens, B., Óskarsdóttir, M. et al. (2024). Special issue on feature engineering editorial. Mach Learn 113, 3917–3928. https://doi.org/10.1007/s10994-021-06042-2

[2] Abdelouahed, S. M., Abla, R., Asmae, E., Abdellah, A. (2024). "Harnessing feature engineering to improve machine learning: A review of different data processing techniques," 2024 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 2024, pp. 1-6, https://ieeexplore.ieee.org/document/10620105

[3] Kulmanov, M., Smaili, F. Z., Gao, X., Hoehndorf, R. (2021). Semantic similarity and machine learning with ontologies. Briefings in Bioinformatics, Volume 22, Issue 4, July 2021, bbaa199. https://doi.org/10.1093/bib/bbaa199

[4] Ghidalia, S., Narsis, O. L., Bertaux, A., Nicolle, C. (2024). Combining Machine Learning and Ontology: A Systematic Literature Review. arXiv preprint arXiv:2401.07744. https://arxiv.org/abs/2401.07744

[5] Faust, K., Bala, S., van Ommeren, R. et al. (2019). Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. Nat Mach Intell 1, 316–321. https://doi.org/10.1038/s42256-019-0068-6

[6] Siddiqui, S., Rehman, M. A., Doudpota, S. M., Waqas, A. (2019). Ontology Driven Feature Engineering for Opinion Mining. IEEE Access, vol. 7, pp. 67392-67401, 2019, doi: 10.1109/ACCESS.2019.2918584. https://ieeexplore.ieee.org/abstract/document/8721082

[7] Sahoo, S.S., Kobow, K., Zhang, J. et al. (2022). Ontology-based feature engineering in machine learning workflows for heterogeneous epilepsy patient records. Sci Rep 12, 19430. https://doi.org/10.1038/s41598-022-23101-3

[8] Kulmanov, M., Smaili, F. Z., Gao, X., Hoehndorf, R. (2020). Machine learning with biomedical ontologies. biorxiv, 2020-05. https://doi.org/10.1101/2020.05.07.082164

[9] Sun, X., Zhang, Y., & Chen, J. (2019). RTPO: A domain knowledge base for robot task planning. *Electronics, 8*(10), 1105. https://doi.org/10.3390/electronics8101105

[10] W3C OWL Working Group. (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)*. World Wide Web Consortium (W3C). https://www.w3.org/TR/owl2-overview/

[11] Cortez, P., & Silva, A.M. (2008). Using data mining to predict secondary school student performance. https://www.semanticscholar.org/paper/Using-data-mining-to-predict-secondary-school-Cortez-Silva/61d468d5254730bbecf822c6b60d7d6595d9889c?sort=relevance&citationIntent=background

ANEXO A: CÓDIGO FUENTE

A través del siguiente link se podrá acceder al código fuente del proyecto:

https://github.com/daniat707/Ontology-Feature-Engineering