

GNN-BIM: Attribute Inference Using Multimodal Embeddings in Building Graphs

No Author Given

No Institute Given

Resumen Building Information Modelling (BIM) environments encode complex relationships between architectural and semantic entities, making them particularly well-suited for a graph-based learning approach. We propose a methodology based on a graph neural network (GNN) to model BIM data as a graph with attributes, fusing semantic, spatial, and topological embeddings. The data originates from federated models of real infrastructure projects, where elements are encoded as nodes with multimodal attributes. By integrating these heterogeneous embeddings, the model learns latent representations of the BIM data. This is addressed as a node-level attribute prediction (classification) problem over the BIM graph.

Keywords: computational geometry, graph theory, graph neural networks, building information modelling.

GNN-BIM: Inferencia de Atributos mediante Embeddings Multimodales en Grafos de Edificación

Resumen. Los entornos de modelado de información de construcción (BIM) codifican relaciones complejas entre entidades arquitectónicas y semánticas, siendo particularmente apropiados para un enfoque de aprendizaje basado en grafos. Se propone una metodología basada en redes neuronales de grafos (GNN) para modelar la información BIM como un grafo con atributos, fusionando *embeddings* semánticos, espaciales y topológicos. La información se origina de modelos federados de proyectos de infraestructura reales, donde los elementos están codificados como nodos con atributos multimodales. Al integrar estos embeddings heterogéneos, el modelo puede aprender representaciones latentes de los datos BIM. Se aborda este aspecto como un problema de predicción de atributos a nivel de nodo (clasificación) sobre el grafo BIM.

Palabras clave: geometría computacional, teoría de grafos, redes neuronales de grafos, modelado de información de edificios.

1. Introducción

El modelado de información de construcción (*BIM*) se había concebido como una oportunidad para consolidar la topología y la semántica de los proyectos en bases de datos colaborativas. Sin embargo, en la industria, la federación de modelos multidisciplinarios mediante el estándar IFC da lugar a un deterioro sistemático de la información. Las diferencias en las prácticas de modelado humano, la duplicación de referencias entre disciplinas y la fragmentación de la semántica estándar dan lugar a inconsistencias y al *Data Drift*, donde los metadatos pueden alterarse de manera impredecible en cada exportación. Tradicionalmente, se ha intentado abordar el enriquecimiento semántico mediante la automatización de conjuntos de reglas basadas en la lógica. Sin embargo, estos métodos no se pueden escalar cuando se aplican en grandes instalaciones industriales, ya que la multitud de posibles excepciones y errores supera rápidamente dichos conjuntos de reglas (Bloch & Sacks, 2018), (Pauwels et al., 2017).

Para superar los retos mencionados, *GNN-BIM* propone el uso de *GNN* para inferir relaciones espaciales implícitas y actuar como un filtro probabilístico de control y garantía de calidad. El modelo utilizado fue *GraphSAGE* (Hamilton et al., 2017).

2. Metodología Propuesta

La metodología se estructura en un proceso automatizado de cuatro fases; Minería de datos espaciales y semánticos, construcción del grafo heterogéneo, entrenamiento inductivo y retroalimentación al modelo *BIM*.

2.1. Extracción de Metadatos y preprocesamiento

Para capturar la topología eficientemente, se extraen las cajas delimitadoras (*AABB*) de cada elemento MEP. Se utiliza *R-Tree* (Guttman, 1984) para calcular colisiones y distancias mínimas de manera eficiente, para más información consultar la sección 6,2 del Anexo. Al mismo tiempo, se extraen los conjuntos de propiedades de cada elemento, donde, para evitar sesgos, se eliminaron los identificadores globales duplicados producto de prácticas de modelado inconsistentes. Estos datos se consolidaron en archivos *Parquet* eliminando atributos con representatividad menor al 1 % mediante un *VarianceThreshold*, acotando de 47,588 a 202 *features* (Consultar sección 6,2 del Anexo para una justificación detallada sobre esta elección).

Finalmente, para evitar *data leakage*, se eliminaron atributos descriptivos (e.g., nombres de sistema) que insinuaran la clasificación.

2.2. Arquitectura, inferencia y métricas

La elección de *GraphSAGE* por sobre *GCN* y *GAT* se fundamenta en su *neighborhood sampling*. Se utilizó el optimizador Adam por su tasa adaptativa,

crucial para grafos desbalanceados. Para mitigar el sesgo de la “cola larga”, se inyectó un tensor de pesos adaptativos en la función $NLLLoss$. El peso $w_c = \frac{N}{C \cdot N_c}$ (donde N es el total de nodos, C la cantidad de clases y N_c el soporte de la clase c) escala asimétricamente los gradientes durante el *backpropagation*, forzando a la red a priorizar patrones minoritarios sobre los elementos masivos. Para más detalles de esta arquitectura ver la sección 6,4 en el Anexo.

La evaluación se realizó mediante validación cruzada estratificada de cinco particiones (80%/20%), que preserva la proporción de cada clase en entrenamiento y prueba para estabilizar las estimaciones frente al desbalance. Se descartó *Accuracy* como métrica principal frente al desbalance encontrado. En su lugar, se adoptaron dos variantes: el *Weighted F1-Score*, que pondera el rendimiento de cada clase según su volumen de aparición (*support*); y el *Macro F1-Score*, que promedia el rendimiento de todas las clases por igual.

3. Caso de Estudio y Resultados

3.1. Dataset y Entorno de Pruebas

Se utilizó el modelo federado de un nivel de un reactor nuclear de investigación a forma de *dataset*. Cada nodo en el grafo representa un elemento MEP, mientras que las aristas homogéneas enlazan aquellos componentes cuyas AABB colisionan físicamente o se encuentran dentro de un umbral de proximidad espacial de 5 cm. Este conjunto presenta el desbalance extremo típico de la disciplina MEP: elementos de conducción masivos frente a componentes arquitectónicos con representación marginal (“cola larga”).

Disponibilidad de Datos: Debido a acuerdos de confidencialidad y seguridad industrial vinculados a la planta nuclear, el *dataset* y el código fuente propietario no pueden ser publicados de forma abierta.

3.2. Métricas de rendimiento y comparación de modelos

El modelo propuesto alcanzó un *Weighted F1-Score* promedio del $95,40\% \pm 0,17\%$ y un *Macro F1-Score* del $70,26\% \pm 2,07\%$ (ver Tabla *benchmark* en material suplementario), confirmando su robustez y estabilidad ante el desbalance. Como se observa en dicha Tabla, *GraphSAGE* supera a los clasificadores clásicos y de grafos en *Macro F1*.

Para validar la resiliencia en escenarios reales de metadatos incompletos, se simuló una caída de información (*Data Drift*) en el conjunto de prueba (Tabla *Data Drift* en Anexo).

La Figura 1 detalla la correlación entre el soporte muestral (n) y la capacidad predictiva. Se observa que la brecha de rendimiento responde a una distribución *long-tail*. Las clases *MEP* mayoritarias superan el 0,85 de F1; su densidad topológica y abundancia consolida la agregación espacial. Al decaer el soporte, la predicción se degrada. No obstante, la pérdida balanceada ponderada impide que la red desestime las clases minoritarias, reteniendo un rendimiento operacional alto para componentes críticos de control que un clasificador tradicional no logra aislar.

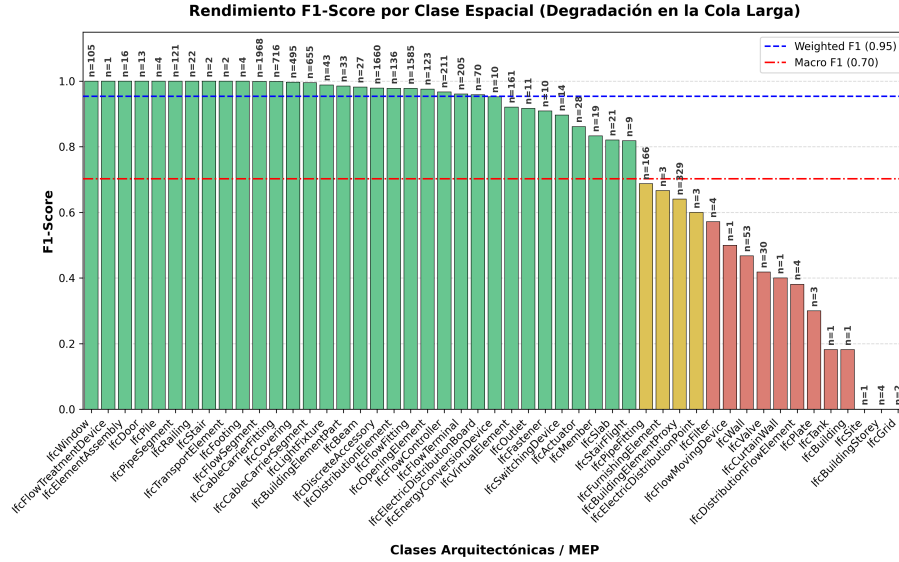


Figura 1. Gráfico de barras F1 Weighted vs. Macro

4. Discusión

El análisis de los resultados permite diagnosticar problemas sistémicos en el ciclo de vida de la información de los modelos BIM a partir de tres factores.

4.1. Resiliencia ante el *Data Drift* y comparación con *baselines*

Ante una pérdida del 50 % de metadatos en el *dataset*, la capacidad predictiva de clasificadores locales colapsa, *Random Forest* a 14,47 % de *Macro F1*. En cambio, el *pipeline* propuesto basado en *GraphSAGE* cae a 53,07 %. A diferencia de los enfoques basados en reglas y *baselines* locales, las capas convolucionales de *GraphSAGE* propagan la información estructural del vecindario *MEP*, logrando inferir con precisión el tipo de componente aun cuando existe *Data Drift*.

4.2. Análisis del desbalance y ambigüedad del IFC

Los análisis demuestran que las clases con bajo soporte y alta ambigüedad espacial, como *IfcValve* y *IfcWall*, caen por debajo del umbral de rendimiento. Al estar físicamente interconectadas con tuberías o elementos estructurales dominantes, sus firmas topológicas locales se solapan, dificultando su clasificación.

A diferencia de trabajos previos en el enriquecimiento semántico de modelos BIM, nuestro modelo asume el desbalance operativo real del modelado coordinado. La inyección de pesos locales adaptativos en la función de pérdida impide

que la *GNN* desestime los *outliers*, permitiendo revelar inconsistencias de gobernanza y etiquetado incorrecto en componentes *MEP* críticos que suelen omitirse en la literatura.

4.3. Validación *Human-in-the-Loop* y Gobernanza de Datos

El uso extendido de clases genéricas como *IfcBuildingElementProxy* en el modelado manual representa una pérdida de calidad semántica en el modelado BIM. Para viabilizar esta herramienta en flujos de producción industrial (*QA/QC*), se propone un enfoque de validación cooperativa *Human-in-the-Loop*. En el mismo la *GNN* provee predicciones, los elementos cuya predicción superen el umbral de confianza se validan automáticamente en modelo federado, mientras que predicciones con baja confianza se marcan para auditoría humana interactiva.

5. Conclusiones y trabajo futuro

Este trabajo demuestra la viabilidad de las redes neuronales de grafos como motor de inferencia para el control de calidad en la industria BIM.

La topología relacional del grafo espacial MEP demostró una tolerancia al fallo crítica frente a la degradación de metadatos locales (*Data Drift*), cumpliendo un rol de agente de auditoría que expone la falta de gobernanza de datos.

Como limitación del estudio, la evaluación se restringió a un único *dataset* correspondiente a un reactor nuclear de investigación complejo. Para trabajos futuros, se plantean tres líneas de investigación:

Como limitación, la evaluación se restringió a un único *dataset* de un reactor nuclear de investigación complejo. Para trabajos futuros, se plantean tres líneas: (1) Generalización de Dominio, evaluando el *pipeline* GNN-BIM en tipologías tradicionales para validar la transferibilidad; (2) Grafos Heterogéneos, implementando arquitecturas para codificar explícitamente la diferencia semántica entre relaciones espaciales; y (3) Inclusión en el *benchmark* de *baselines* basadas en reglas heurísticas.

6. Anexos

6.1. Glosario

- **BIM:** *Building Information Modelling*. Metodología de trabajo colaborativo para la creación y gestión de un proyecto de construcción, centralizando toda la información en un modelo digital multidimensional.
- **MEP:** *Mechanical, Electrical, and Plumbing*. Disciplinas referidas a las instalaciones mecánicas, eléctricas y sanitarias de una infraestructura.
- **BEP:** *BIM Execution Plan*. Documento rector que define cómo se llevará a cabo el modelado de información, estableciendo roles, estándares de datos y flujos de trabajo para el proyecto.
- **IFC:** *Industry Foundation Classes*. Formato de datos estándar, abierto y neutral (ISO 16739) utilizado para el intercambio interoperable de modelos BIM entre distintas plataformas de software.
- **Modelo Federado:** Ensamblaje de distintos modelos BIM multidisciplinares en un único entorno coordinado, utilizado principalmente para la detección de interferencias espaciales y la auditoría del proyecto.
- **AABB:** *Axis-Aligned Bounding Boxes*. Volumen rectangular paralelo a los ejes coordenados que engloba la geometría tridimensional de un objeto para simplificar el cálculo de colisiones.
- **QA/QC:** *Quality Assurance / Quality Control*. En el contexto BIM, abarca los procesos probabilísticos y heurísticos para garantizar que el modelo cumple con los estándares de información requeridos y está libre de colisiones físicas o metadatos erróneos.

6.2. Metadatos y preprocesamiento

Justificación de uso *R-Tree* El uso de árboles R (*R-Tree*) se justifica por su eficiencia algorítmica en la indexación espacial multidimensional. Dado que contrastar las colisiones entre cajas delimitadoras (*AABB*) por fuerza bruta requiere una complejidad computacional de $O(N^2)$, el *R-Tree* agrupa los elementos por proximidad jerárquica. Esto reduce drásticamente el tiempo de búsqueda a $O(\log N)$ para consultas de intersección y distancias mínimas, permitiendo extraer la topología y construir el grafo espacial MEP a gran escala de forma viable.

Justificación uso de archivos *Parquet* La naturaleza binaria de estos permite consumir las matrices directamente en memoria (RAM/VRAM), maximizando el *throughput* y la escalabilidad computacional.

Funcionamiento *Variance Threshold* Para mitigar la dimensionalidad y eliminar ruido estadístico se aplica un filtro de varianza, *VarianceThreshold*, con un umbral de 0,01. Esta técnica remueve columnas con representatividad menor al 1 % sin alterar la variabilidad estructural del conjunto de datos.

6.3. Diferencias entre sistemas expertos e inteligentes

Tradicionalmente, el enriquecimiento semántico y el control de calidad en la industria BIM se han abordado mediante sistemas expertos, consistentes en conjuntos de reglas lógicas o heurísticas codificadas manualmente (basadas típicamente en atributos categóricos). A diferencia de los sistemas inteligentes basados en aprendizaje automático (como la arquitectura *GraphSAGE* propuesta), que infieren patrones subyacentes a partir de la topología y el contexto geométrico del grafo, los enfoques basados en reglas presentan dos debilidades operativas críticas:

1) *Costo de Ingeniería*: Requieren el diseño, mantenimiento y actualización de reglas complejas hechas a mano que no escalan ni generalizan ante nuevos proyectos o convenciones de modelado. 2) *Fragilidad*: Dependen estrictamente de la integridad de los datos de entrada. Si la nomenclatura de nombres se corrompe, los atributos están vacíos o existe *Data Drift*, el sistema de reglas colapsa de forma frágil, reduciendo su precisión algorítmica al 0 %.

6.4. Arquitectura

Justificación de *GraphSAGE* Se eligió este método por sobre *GCN* y *GAT* ya que su operador de concatenación preserva el *embedding* del nodo de forma diferenciada del mensaje del vecindario, mientras que su agregación sobre un muestreo uniforme acotado previene el sobre-suavizado (*over-smoothing*) y la inestabilidad en la proyección de *weights* que sufren *GCN* y *GAT* ante la dispersión de grados y el ruido topológico espacial del estándar IFC.

Detalles de arquitectura *GraphSAGE* *neighborhood sampling* de tamaño fijo ($k = [15, 10]$, $L = 2$), garantizando un costo predecible $O(k^L)$ independiente de la densidad del grafo (Wu et al., 2020)

6.5. Tabla Benchmark

Análisis Tabla *Benchmark* *Random Forest* y *MLP* reportan un *Weighted F1* ligeramente superior (96 %), su rendimiento *Macro F1* cae por debajo del 67,5 %, revelando ineficiencia en las clases minoritarias. Por otro lado, las arquitecturas basadas en grafos *GCN* y *GAT* muestran una degradación en sus métricas. Esto se debe a que estos modelos propagan mensajes de manera global, sin discriminar la heterogeneidad y escasez de metadatos en un grafo inherentemente desbalanceado, sufriendo *over-smoothing* y *overfitting*.

Análisis Tabla *Data Drift* Ante una pérdida del 50 % de atributos en test, el *Macro F1* de *Random Forest* colapsa un 78,1 % relativo (cayendo a 14,47 %), mientras que *GraphSAGE* retiene un robusto 53,07 % (reducción de solo el 24,5 %). Ante una pérdida masiva del 80 %, *GraphSAGE* supera por cinco veces a *Random Forest* (16,66 % vs. 3,38 %). Esto evidencia que la *GNN* compensa las propiedades locales nulas agregando el contexto topológico y geométrico de los componentes conectados en el grafo *MEP*.

Cuadro 1. Resultados del benchmark comparativo (5-Fold Stratified CV).

Modelo	Weighted F1 (%)	Macro F1 (%)
Random Forest	96, 11 $\pm$ 0, 20	67, 44 $\pm$ 1, 69
MLP (Red Neuronal)	96, 88 $\pm$ 0, 13	67, 31 $\pm$ 1, 06
GCN (Grafo Homogéneo)	43, 38 $\pm$ 1, 41	26, 67 $\pm$ 1, 93
GAT (Grafo con Atención)	47, 03 $\pm$ 1, 37	33, 02 $\pm$ 1, 24
GraphSAGE (Propuesto)	95, 40 $\pm$ 0, 17	70, 26 $\pm$ 2, 07

Cuadro 2. Resiliencia ante pérdida de atributos (Data Drift) en Test.

Modelo	Macro F1 (%)	Drift 50 % (%)	Drift 80 % (%)
Random Forest	66, 22	14, 47	3, 38
GraphSAGE (Propuesto)	70, 26	53, 07	16, 66

Referencias

- Bloch, T., & Sacks, R. (2018). Comparing machine learning and rule-based inferencing for semantic enrichment of BIM models. *Automation in Construction*, 91, 256-272.
- Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, 47-57.
- Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.
- Pauwels, P., Zhang, S., & Lee, Y.-C. (2017). Semantic web technologies in AEC industry: A literature overview. *Automation in Construction*, 73, 145-165.
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4-24.