Redes neuronales de grafos heterogéneos: aplicación en la medicina

Josué Santiago Cano Pulgarín

Trabajo de grado presentado para optar al título de Ingeniero de Sistemas

Asesor
Raul Ramos Pollan, Doctor (PhD) en Ingeniería Informática

Universidad de Antioquia
Facultad de Ingeniería
Ingeniería de Sistemas
Medellín, Antioquia, Colombia
2023

Cita	Cano Pulgarín [1]
Referencia Estilo IEEE (2020)	[1] J. Cano Pulgarín y R. Ramos Pollan, “Redes Neuronales de grafos heterogéneos: aplicación en la medicina”, Trabajo de grado profesional, Ingeniería de Sistemas, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2023.

CONTENTS

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

En este estudio, se investigó el uso de redes neuronales de grafos heterogéneas (HetGNN) para el análisis de datos médicos estructurados y la predicción de readmisión hospitalaria en pacientes con diabetes. Se propuso una metodología que implicaba la conversión de la base de datos tabular a una representación de grafo, y se comparó el rendimiento de los modelos de HetGNN con los modelos tradicionales de machine learning utilizados en la línea base. Se realizaron experimentos utilizando diferentes técnicas de resampling y se evaluaron métricas de evaluación, como el AUC y el recall. Además, se exploró la influencia de la complejidad del grafo en el rendimiento del modelo y se ajustaron parámetros clave, como el learning rate y el número de epochs.

Los resultados obtenidos mostraron que los modelos de HetGNN, especialmente cuando se utilizaban técnicas de resampling como undersampling, demostraron un rendimiento prometedor en términos de precisión y capacidad de discriminación entre clases. Se observó un desempeño levemente inferior en la capacidad predictiva en comparación con la línea base, pero se destacó el potencial de las HetGNN para capturar las relaciones complejas entre las variables médicas en forma de grafo. El análisis de datos médicos estructurados a través de representaciones de grafos permitió una nueva perspectiva en la comprensión de los factores que influyen en la readmisión hospitalaria en pacientes con diabetes.

Palabras clave — **Redes neuronales de grafos, Medicina, Inteligencia artificial en Medicina, GNN, AI.**

ABSTRACT

This study investigated the use of heterogeneous graph neural networks (HetGNN) to analyze structured medical data and predict hospital readmission in patients with diabetes. A methodology was proposed that involved converting the tabular database into a graph representation, and the performance of HetGNN models was compared to traditional machine learning models used in the baseline. Experiments were conducted using different resampling techniques, and evaluation metrics such as AUC and recall were assessed. Furthermore, the influence of graph complexity on model performance was explored, and critical parameters such as learning rate and number of epochs were adjusted.

The results demonstrated that HetGNN models, particularly when combined with undersampling techniques, exhibited promising accuracy and class discrimination performance. Although slightly lower predictive capacity was observed compared to the baseline, HetGNN showed potential in capturing the complex relationships among medical variables as a graph. Analyzing structured medical data through graph representations provided a new perspective in understanding the factors influencing hospital readmission in patients with diabetes.

Keywords — **Scientific article, review article, research, citation styles. Graph Neural Networks, Medicine, Artificial Intelligence in Medicine, GNN, AI.**

1. INTRODUCCIÓN

La aplicación de la tecnología en el campo de la medicina ha traído consigo avances notables que han salvado innumerables vidas. La inteligencia artificial (IA) ha surgido como una herramienta poderosa, aportando modelos de predicción que ayudan a los profesionales de la salud a tomar decisiones informadas y prevenir enfermedades [1]. Estas aplicaciones han abarcado diversos ámbitos, desde el aprendizaje supervisado utilizando datos estructurados hasta la utilización de datos no estructurados como imágenes o texto. Sin embargo, al analizar datos médicos estructurados, surgen desafíos inherentes debido a la limitada capacidad para capturar relaciones implícitas. Descubrir estas relaciones y aprovecharlas para mejorar el análisis podría proporcionar conocimientos valiosos para mejorar el rendimiento del modelado.

En los últimos años, ha surgido un creciente interés en el concepto de representar datos médicos en forma de grafos, abriendo así el camino para la utilización de redes neuronales de grafos en este campo [2]. Los grafos ofrecen un marco más expresivo y poderoso para manejar conceptos complejos como relaciones e interacciones. A diferencia de las representaciones tabulares tradicionales, los grafos brindan una representación visual intuitiva que captura las intrincadas interdependencias entre las variables médicas. Al utilizar la naturaleza relacional de los datos médicos, los grafos ofrecen una base natural para analizar las relaciones dentro de un marco contextual y tienen el potencial de simplificar problemas complejos o verlos desde diferentes perspectivas, lo que conduce a análisis más completos e ilustrativos.

Lamentablemente, muchas bases de datos médicas existentes aún se estructuran en formato tabular, lo que limita la aplicación de modelos basados en grafos, que requieren el desarrollo de una metodología de conversión que transforme de manera efectiva conjuntos de datos tabulares en representaciones de grafos, permitiendo la aplicación de redes neuronales de grafos heterogéneas (HetGNN) y la comparación con otros tipos de modelos.

De acuerdo con lo anterior, la presente investigación tiene como objetivo comparar el perfomance de la aplicación de arquitecturas de redes neuronales heterogéneas de grafos con modelos tradicionales de machine learning en bases de datos estructuradas.

En las secciones siguientes, esta tesis se adentrará en una revisión exhaustiva de la literatura, discutiendo enfoques existentes para el análisis de datos médicos, las limitaciones de las representaciones tabulares y las tendencias emergentes en la utilización de estructuras basadas en grafos y redes neuronales de grafos. Se presentará en detalle la metodología propuesta para la conversión de datos tabulares a grafos, incluyendo los pasos involucrados, la elección de algoritmos de construcción de grafos y técnicas de preprocesamiento de datos. También se describirá la configuración experimental y las métricas de evaluación, seguidas de un análisis y discusión exhaustivos de los resultados obtenidos. Además, se abordarán los desafíos y las limitaciones potenciales del enfoque propuesto, junto con recomendaciones para futuras direcciones de investigación. Por último, se discutirán las conclusiones extraídas de esta investigación, resaltando las contribuciones, limitaciones y posibles implicaciones en el campo del análisis de datos médicos y la modelización basada en grafos.

1.1. PLANTEAMIENTO DEL PROBLEMA

En la era actual de la informática y el análisis de datos, la aplicación de técnicas de machine learning se ha convertido en un pilar fundamental para extraer conocimiento y realizar predicciones en diversos campos. Esto cobra especial relevancia en el área de la medicina, en el cual modelos predictivos óptimos pueden contribuir de manera significativa a mejorar la calidad de vida de las personas, e incluso, a salvar vidas. En esta área, la mayor parte de información se encuentra como bases de datos tradicionales en formato tabular, lo que puede limitar la representación y el análisis adecuado de las relaciones complejas entre entidades y variables; por ello, múltiples investigaciones de IA en el campo de la medicina se han centrado en el análisis de otro tipo de datos, entre ellos, imágenes, texto y audio, dejando de lado una gran cantidad de información con alto potencial.

Lo anterior, plantea la necesidad de mejorar los resultados obtenidos a partir de bases de datos tabulares, siendo una opción viable la conversión de estas en grafos y la posterior utilización de redes neuronales de grafos para el análisis y la predicción. Esta metodología permitiría aprovechar la estructura de los grafos para capturar de manera más eficiente las relaciones entre los datos, lo que podría resultar en un aumento significativo en la precisión y la capacidad de generalización de los modelos de machine learning. Sin embargo, hasta el momento, no se han encontrado estudios en los que se realice la conversión de bases de datos tabulares a grafos en ningún campo, o se evalúe la efectividad de los grafos frente a los métodos tradicionales de maching learning, por lo que este asunto resulta un vacío de investigación que debe ser abordado.

Adicionalmente, las estructuras de grafos se han abordado en la mayoría de investigacines desde la perspectiva de grafos homogéneos, dejando un vacío en la comprensión y el aprovechamiento de la heterogeneidad presente en los grafos, lo cual implica mayor complejidad pero ofrece mayores posibilidades de obtención de información, ya que, específicamente en medicina, la mayoría de las bases de datos presentan esta característica.

De acuerdo con lo anterior, se plantea el problema de investigación con la siguiente pregunta: ¿Cómo afecta la conversión de una base de datos tabular a un grafo y la utilización de redes neuronales de grafos heterogéneos el rendimiento de los modelos de machine learning aplicados a una base de datos?

1.2. JUSTIFICACIÓN

El presente proyecto de investigación representa la primera intervención sobre bases tabulares, realizando su conversión a grafos y evaluando la efectividad de estos con respecto a los modelos tradicionales de machine learning. Este análisis permitirá comprender en qué medida las HetGNN pueden capturar y utilizar eficientemente la información de los diferentes tipos de nodos y relaciones en los grafos heterogéneos.

La relevancia de esta investigación radica en su contribución al avance del análisis de datos estructurados. Al abordar la heterogeneidad de los grafos, se ampliará el conocimiento y las capacidades en el campo del análisis de datos estructurados, permitiendo su aplicación en áreas que generalmente recolectan la información de forma estructurada, como la medicina, la biología y las redes sociales, entre otras. Además, esta investigación contribuirá a identificar las ventajas y limitaciones de las arquitecturas de redes neuronales heterogéneas de grafos en comparación con los modelos tradicionales.

La aplicabilidad de los resultados obtenidos en este estudio es otro aspecto relevante. La resolución de problemas complejos en diversas áreas se beneficiará de los hallazgos de esta investigación. La evaluación del rendimiento de las HetGNN en bases de datos estructuradas brindará información práctica para enfrentar problemas reales y mejorar la toma de decisiones.

2. TRABAJO RELACIONADO

2.1. Redes neuronales de grafos

En los últimos años, las redes neuronales de grafos han ganado atención significativa como un enfoque prometedor para abordar la clasificación de nodos en datos de grafos. Estas redes han demostrado su eficacia para modelar las relaciones complejas y estructurales presentes en los grafos, mejorando así la precisión de la clasificación. A continuación, se presentan algunos trabajos clave en este campo, junto con sus respectivas referencias bibliográficas.

Una de las primeras propuestas en este ámbito es el modelo Graph Convolutional Network (GCN) [3]. Esta arquitectura utiliza operaciones de convolución en grafos para aprender representaciones de nodos, capturando la información estructural y relacional del grafo. El GCN ha sido ampliamente adoptado en diversas tareas de clasificación de nodos en grafos.

Otro enfoque importante es el GraphSAGE (Graph Sample and Aggregated) [4], que utiliza técnicas de muestreo y agregación para generar representaciones de nodos en grafos grandes y heterogéneos. Esta técnica permite el aprendizaje eficiente de características para la clasificación de nodos en grafos de gran escala.

Además, se ha propuesto el modelo Graph Attention Network (GAT) [4], que utiliza mecanismos de atención para capturar las relaciones de importancia entre los nodos vecinos durante el procesamiento de la información. Esta arquitectura permite una mayor flexibilidad en la captura de información relevante en diferentes vecindarios de nodos.

Otros enfoques, como GraphSAGE con estructuras de grafo inductivas [5], Graph Isomorphism Networks (GIN) [6], y Graph Neural Networks (GNNs) recurrentes [7], también han demostrado su eficacia en la clasificación de nodos en diferentes contextos.

2.2. Redes neuronales heterogéneas de grafos

En los últimos años, ha habido un creciente interés en el desarrollo de redes neuronales heterogéneas de grafos para abordar el problema de la clasificación de nodos en datos de grafos heterogéneos. Estas redes neuronales han demostrado ser efectivas para modelar la complejidad y la diversidad de los datos en grafos heterogéneos, mejorando así el rendimiento de la clasificación de nodos. A continuación, se presentan algunos trabajos relevantes en este campo, junto con sus respectivas referencias bibliográficas.

Una de las primeras propuestas en este ámbito es el modelo HAN (Hierarchical Attention Network) [5]. Este enfoque propone una red neuronal heterogénea de grafos basada en una arquitectura jerárquica. Utiliza mecanismos de atención separados para capturar información

detallada de los nodos y a nivel semántico, permitiendo un aprendizaje de representaciones efectivo.

Otra propuesta importante es el modelo HGT (Heterogeneous Graph Transformer) [4]. Este modelo adopta una arquitectura basada en transformers para modelar las dependencias entre diferentes tipos de nodos en un grafo heterogéneo. Utiliza mecanismos de atención con múltiples cabezas y codificaciones posicionales para capturar información contextual rica, lo que mejora el rendimiento de la clasificación de nodos.

Además, se ha explorado la combinación de mecanismos de atención y unidades recurrentes con compuertas (GRU) en el modelo HAN-GRU (Heterogeneous Graph Attention Network with Gated Recurrent Units) [5]. Este enfoque permite capturar dependencias secuenciales y atender a los nodos vecinos relevantes durante el proceso de propagación de mensajes.

Otro modelo relevante es HetGNN (Heterogeneous Graph Neural Network) [6], el cual se ha diseñado específicamente para grafos heterogéneos. Incorpora operaciones de agregación y transformación que tienen en cuenta la heterogeneidad de los nodos y las aristas, permitiendo una clasificación precisa de los nodos.

Por último, el modelo MMGCN (Multi-modal Graph Convolutional Network) [7] se centra en la clasificación de nodos en grafos heterogéneos con múltiples modalidades de información. Combina redes convolucionales de grafos específicas para cada modalidad y utiliza capas de fusión para integrar la información de diferentes modalidades, mejorando así el rendimiento de la clasificación.

Estos trabajos representan solo una muestra de los avances en el campo de las redes neuronales heterogéneas de grafos para la clasificación de nodos. Cada uno de ellos ha contribuido al desarrollo de metodologías y técnicas innovadoras que permiten abordar los desafíos de la heterogeneidad en los datos de grafos y mejorar el rendimiento de la clasificación de nodos en grafos heterogéneos.

3. MÉTODO PROPUESTO

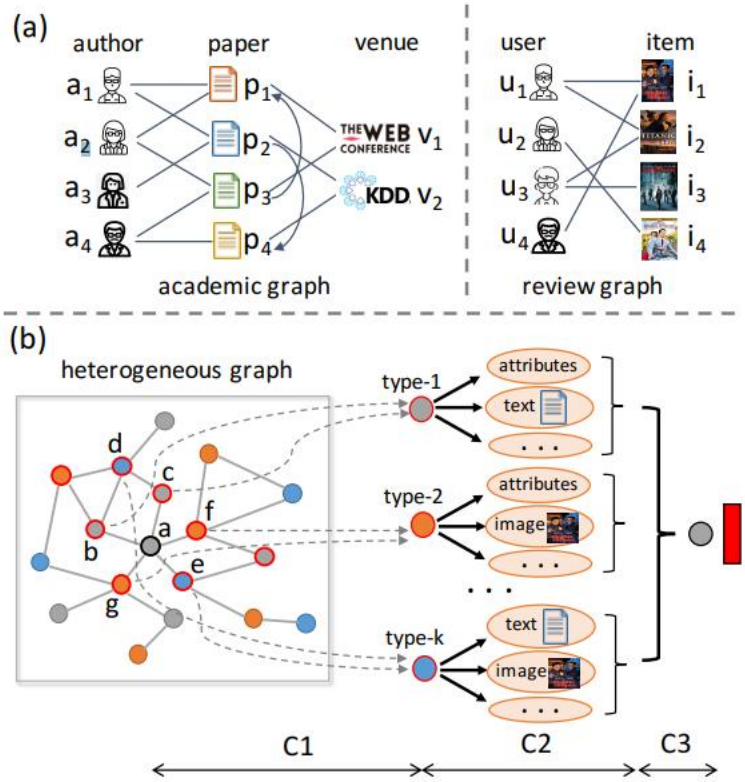
3.1. Definición de Grafos

Un grafo es una estructura de datos que consta de dos componentes: vértices y aristas. Se utiliza como una estructura matemática para analizar la relación de pares entre objetos y entidades. Típicamente, un grafo se define como G = (V, E), donde V es un conjunto de nodos y E son las aristas entre ellos. La idea intuitiva que subraya el enfoque propuesto es que los nodos en un grafo representan objetos o conceptos, y los bordes representan sus relaciones. Cada concepto se define naturalmente por sus características y conceptos relacionados [8].

Teniendo esto en cuenta, a cada nodo obtiene un estado (x) para representar su concepto. Podemos usar el estado del nodo (x) para producir una salida (o), es decir, una decisión sobre el concepto. El estado final (xn) del nodo normalmente se denomina "incrustación de nodos". La tarea de todos los Graph Neural Network (GNN) es determinar la "incrustación de nodos" de cada nodo, observando la información de sus nodos vecinos.

Muchas aplicaciones del mundo real consideran mas de un tipo de nodo o de relación, cosa que los GNN no permiten, por lo que se desarrollaron los HetGNN [9] que permiten la utilización de diferentes tipos de nodos y relaciones.

Figura 1



(a) Ejemplos de HetG: un gráfico académico y un gráfico de revisión. (b) Desafíos de la red neuronal de grafos para HetG: C1- muestreo de vecinos heterogéneos (para el nodo a, en este caso, los colores del nodo denotan diferentes tipos); C2 - codificación de contenidos heterogéneos; C3 – agregación de vecinos heterogéneos [9].

3.2. Modelo HeteGCN

Una de las aproximaciones más comunes son los Heterogeneus Graph Convolutional Network (GCN)[8]. Se analizará esta arquitectura como referencia del funcionamiento general de un HetGNN. Un concepto fundamental es que cada nodo está conectado consigo mismo, que es lo mismo que $(v, v)_{\in}E$, el resto de las conexiones vamos a definir a X como una matriz que contiene tonos lo nodos (n) son sus características, formalmente definido como $X_{\in}R^{n_x m}$ siendo m la dimensión del vector de características. Con esta información se puede definir las GCN [9] que es una red neuronal multicapa que opera directamente sobre un grafo e induce vectores de incrustación de nodos en función de las propiedades de sus vecindarios, matemáticamente consideramos una GCN multicapa con la siguiente regla de propagación por capas:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \quad [8]$$

Siendo, $\tilde{A} = A + I_N$ la matriz de adyacencia del grafo no dirigido G con auto conexiones añadidas; IN es la matriz identidad; $\widetilde{D}$ es la matriz de grados de la matriz $\tilde{A}$ y está definida como $\widetilde{D}_{ii} = \sum_j \tilde{A}_{ij}\, A + I_N$; $W^{(l)}$ es una matriz de peso entrenable específica de la capa l; $H^{(l)}{}_{\epsilon} R^{N_x D}$ es la matriz de activación en la lth capa siendo $H^{(0)} = X$, finalmente $\sigma(\cdot)$ denota la función de activación. La salida de la última capa del modelo pasa por la función softmax, la cual está definida como: $softmax(x_i) = \frac{1}{Z}\exp(x_i)$ con $Z = \sum_i \exp(x_i)$ y siendo aplicada en cuanto a filas [9]. Este modelo de red neuronal al igual que la mayoría de las redes utiliza como función de perdida el cross-entropy error sobre todas las etiquetas de la siguiente manera:

$$\mathcal{L} = -\sum_{l \in \mathcal{Y}_L} \sum_{f=1}^{F} Y_{lf} \ln Z_{lf}$$

Aquí, F son los mapas de entidades en la capa de salida; YL es el conjunto de índices de nodos que tienen etiquetas y Z es la aplicación del softmax.

3.3. Metodología Propuesta

3.3.1. Línea Base

En primer lugar, se llevó a cabo el desarrollo de una línea base que serviría como punto de comparación con los modelos de grafos propuestos. Para ello, se decidió implementar la metodología descrita en [referencia]. Esta metodología se caracteriza por la creación de 36 agrupaciones de enfermedades, las cuales se generan a partir de los tres diagnósticos presentes en la base de datos. Además, se limitó el uso de registros a aquellos que involucraran medicamentos para la diabetes. Si bien la base de datos original tenía tres categorías para predecir (No readmisión, readmisión < 30 días, readmisión > 30 días), se simplificó a una clasificación binaria: No readmisión y Readmisión. Los modelos seleccionados para la línea base fueron XGBoost, Random Forest y Neural Network, a los cuales se les aplicó un ajuste de hyperparámetros para obtener los mejores resultados.

3.3.2. Conversión base tabular a grafo

Una vez obtenida la línea base, se procedió a la conversión de los datos a formato de grafo. Se consideraron como nodos a las personas, las enfermedades diagnosticadas y los medicamentos aplicados, y se establecieron relaciones entre los nodos de persona-medicamento y persona-enfermedad. La creación del grafo se realizó mediante la construcción de matrices de adyacencia para ambos tipos de relaciones, siendo las filas las personas y las enfermedades y medicamentos las columnas en cada caso. Con estas matrices de adyacencia, se generó un grafo heterogéneo utilizando la librería DGL [16]. El objetivo de este grafo se centró en la clasificación, enfocándose en el nodo de persona. Además, las características restantes se agregaron al nodo de persona con el fin de utilizarlas en el proceso de entrenamiento. Dado el bajo número de conexiones utilizando las enfermedades en las agrupaciones utilizadas en la línea base, se decidió crear una segunda versión del grafo. En esta versión, se establecieron conexiones a los tres diagnósticos, dando lugar a 32 diagnósticos diferentes, y cada nodo de persona pudo estar conectado hasta con tres diagnósticos, lo que generó un grafo con mayor complejidad.

3.3.3. Arquitectura del modelo

Se desarrolló una arquitectura de red neuronal de grafos heterogéneos. Se utilizaron dos capas de HeteroGraphConv, seguidas de tres capas de Linear. Para los nodos de persona, se emplearon las características mencionadas anteriormente, mientras que para los otros dos tipos de nodos se utilizaron las características embebidas de la red. También se planteó un caso experimental en el cual solo se utilizaron las características embebidas de la red para todos los tipos de nodo. Además, se incorporaron nodos adicionales, como raza, edad y género, que tenían la capacidad de establecer relaciones con el nodo de persona, aumentando así la complejidad del grafo.

3.3.4. Resampling

Debido al desbalance significativo entre las categorías a predecir, se emplearon dos métodos para abordar este problema. En primer lugar, se utilizaron pesos en la función de costo para otorgar un mayor castigo al error de la clase minoritaria. En segundo lugar, se aplicaron métodos de resampling en la base de datos tabular. Se dividió la base de datos original en una sección de entrenamiento, a la cual se le aplicó el método de resampling, y una segunda sección de prueba, que se convirtió directamente en grafo para evaluar la eficacia del modelo. Los métodos de resampling utilizados incluyeron undersampling mediante Cluster Centroids, RandomUnderSampler y TomekLinks, oversampling con SMOTE, y una combinación de ambos mediante SMOTETomek [12]. Además, se realizaron pruebas con diferentes tipos de learning rate y epochs, los cuales mostraron tener un impacto significativo en el entrenamiento de los modelos. Finalmente, se siguió la metodología de entrenamiento descrita por la librería DGL [referencia].

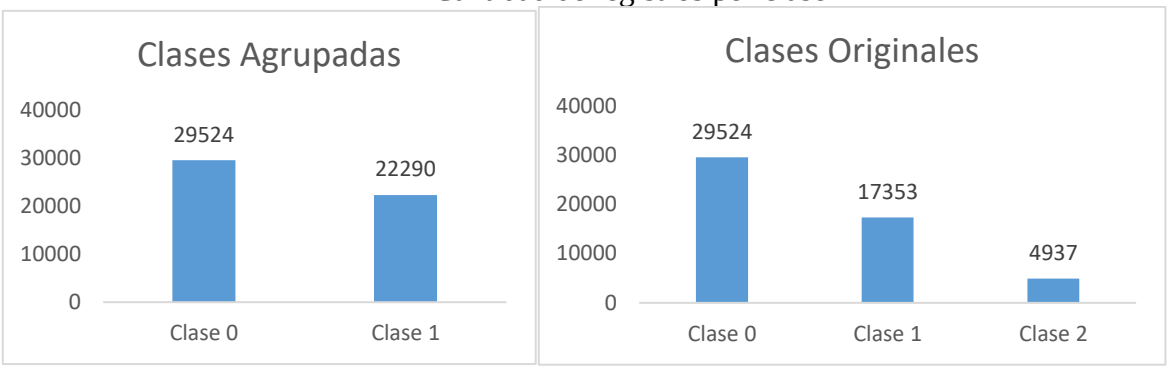
4. EXPERIMENTOS Y ANÁLISIS

4.1. Bases de datos y métricas de evaluación

El presente proyecto se basó en una base de datos de readmisión hospitalaria de diabetes ampliamente utilizada en el desarrollo de modelos tradicionales de machine learning. Esta base de datos contiene información sociodemográfica de los pacientes además de información relacionada con su estadía en el centro asistencial tales como enfermedades diagnosticadas (tres diagnósticos) medicamentos aplicados, tiempo de estadía, entre otros. La base de datos contiene tres etiquetas a predecir: no readmisión, readmisión < 30 días y readmisión > 30 días, se simplifico a una predicción binaria agrupando las etiquetas de readmisión quedando como etiquetas: No readmisión y Readmisión. Se presenta un fuerte desbalance entre las clases a predecir como se puede observar en (Figura 1). Debido a el desbalance de la base de datos se utilizará como métrica principal el AUC y como secundaria el Recall esto debido a que para esta base de datos la clase positiva es de extremo valor

Figura 2

Cantidad de registros por clase



4.1.1. Base de datos principal

Se decidió implementar la metodología descrita en [referencia]. Esta metodología se caracteriza por la creación de 36 agrupaciones de enfermedades, estas agrupaciones fueron definidas por expertos y constan de combinaciones de dos diagnósticos, de los tres informados al paciente durante su estadía. Adicionalmente se procedió con una agrupación de los medicamentos debido a la presencia de medicación con baja frecuencia de aparición. Se limitó el uso de registros a aquellos que involucraran medicamentos para la diabetes además de usar únicamente registros únicos. Se eliminaron características con un porcentaje de valores perdidos mayor al 2% y cualquier registro restante con valores perdidos fue eliminado.

Esta base de datos convertida a grafos creara 36 nodos para enfermedades y cada nodo persona estará conectado con un nodo enfermedad.

El diseño planteado en [11] propone una categorización sobre tres etiquetas a predecir: no readmisión, readmisión < 30 días y readmisión > 30 días. Para efectos de la experimentación en grafos y por ende en esta línea base se simplifico a una predicción binaria agrupando las etiquetas de readmisión quedando como etiquetas: No readmisión y Readmisión.

4.1.2. Variación de la base principal

Debido a la baja frecuencia de aparición que tienen algunas clasificaciones de enfermedades se decidió hacer una variación en este apartado en aras de mejorar el desempeño de la red de grafos. Se decidió para esta variación no clasificar las enfermedades, por el contrario, se usaron las 32 principales y los tres diagnósticos.

Esto llevado a los grafos significa que cada nodo persona está conectado hasta a tres enfermedades, una por diagnostico generando así una mayor cantidad de conexiones en el grafo.

4.2. Configuración experimental

Línea base: Se llevó a cabo el desarrollo de una línea base que serviría como punto de comparación con los modelos de grafos propuestos. Para ello, los modelos seleccionados para la línea base fueron XGBoost [13], Random Forest [14] y Neural Network [15], a los cuales se les aplicó un ajuste de hiperparámetros para obtener los mejores resultados.

Grafo base: Se diseño un grafo con los tres nodos base (Persona, Medicamento y Enfermedades), este grafo se desarrollará para ambas variaciones de base de datos propuestas.

Grafo complejo: Se diseñaron grafos con los tres nodos base (Persona, Medicamento y Enfermedades) más la adhesión de 3 tipos de nodos extra. Se creará un grafo con los tres nodos base más el nodo raza, otro con los anteriores y el nodo género y finalmente con la adhseion del nodo edad, estos grafos se desarrollarán para ambas variaciones de base de datos propuestas.

Tabla 1
RESULTADOS LINEA BASE

Resampling	Train		Test	
	AUC	**Recall**	**AUC**	**Recall**
XGBoost	**0.714**	**0.710**	0.587	0.570
Random Forest	0.652	0.67	**0.600**	**0.600**
Neural Network	0.622	0.650	0.597	0.580

4.3. Primer Acercamiento

Los primeros experimentos realizados sobre el grafo base utilizando las características de la base en el nodo persona no obtuvieron resultados significativos (AUC: 0.5, Recall: 0), lo cual es contrario a la hipótesis planteada por el presente trabajo. Por ello se puso foco en el profundo desbalance de las clases. La experimentación con diferentes técnicas de resampling utilizando la misma estructura obtuvo los mismos resultados. Se realizaron experimentos con los grafos complejos obteniendo resultados inversos a los iniciales, pero aun sin ser significativos (AUC: 0.5, Recall: 1).

El factor primordial que se descubrió fue la utilización única de los embedings del modelo de grafo como características de cada tipo de nodo, eliminando así el uso de las características de labase en el nodo persona. Con los primeros modelos de clasificación con capacidad de discriminación entre clases mediante el uso de diferentes estructuras de nodos, se estableció una base para la mejora mediante el resampling. Las técnicas de resampling tenían como requisito igualar las muestras del conjunto de entrenamiento, para así quitar el bias del desbalance. Por tal motivo los resultados presentados a partir de este punto utilizo las características embebidas de la red de grafos como características de entrenamiento.

4.4. Variaciones de bases de datos

En primer lugar, la creación de dos versiones de las bases de datos permitió probar la importancia de la variabilidad en las diferentes conexiones que el grafo debe tener. Se compararon las dos versiones de bases de datos que se propusieron, se realizó esta comparación en base a la técnica de resampling, para esto el tipo de grafo utilizado fue el base (Tabla 1). Con esta pequeña modificación los resultados base obtienen una mejora. La implementación de las técnicas de resampling mostró

una mejora en la capacidad predictiva de los modelos y más importante aún, la capacidad de discriminación sobre la clase positiva.

Se evidencio una mayor variabilidad del AUC en la base 1. Se evidenciaron valores de recall especialmente bajos en ambos tipos de base. Solo dos metodos de resampling obtuvieron valores de recall mayores a 0.5 los cuales fueron Random Undersampling y Cluster Centroids para la base 1 y 2 respectivamente. El mejor resultado fue un Random Undersampling para la base 1 tanto para AUC como Recall.

Tabla 2
COMPARACION ENTRE VARIACIONES DE LA BASE POR TECNICA DE RESAMPLING

Resampling	Base 1		Base 2	
	AUC	Recall	AUC	Recall
Original	0.525	0.402	0.532	0.212
Cluster Centroids	0.525	0.439	0.528	0.598
RandomUnderSampler	0.590	0.590	0.544	0.509
TomekLinks	0.508	0.054	0.541	0.345
SMOTE	0.524	0.554	0.525	0.338
SMOTETomek	0.527	0.445	0.523	0.353

4.4. Grafos complejos

La diferencia entre los resultados de cada tipo de base junto con el no uso de las características del nodo persona permitió explorar la mejora del performance mediante la complejizacion del grafo. Para esto se partió de los tres nodos principales y se fue agregando un nuevo tipo hasta llegar a los 6. En conjunto de entrenamiento se notó una clara tendencia de a mayor complejidad del grafo mejor desempeño, sin importar que método de resampling se utilizará. Aunque esto no se translado completamente al conjunto se evidencia que tres de los cinco métodos si siguen está tendencia. La implementación de los métodos de resampling deja al random undersampling como el método más eficiente, aunque este método es el menos elaborado de los cinco utilizados nos permite evidenciar que la hipótesis planteada ejecutada sobre un conjunto de datos balanceado podría obtener un desempeño favorable.

Figura 3

DESEMPEÑO POR NUMERO DE TIPOS DE NODOS (TRAIN)

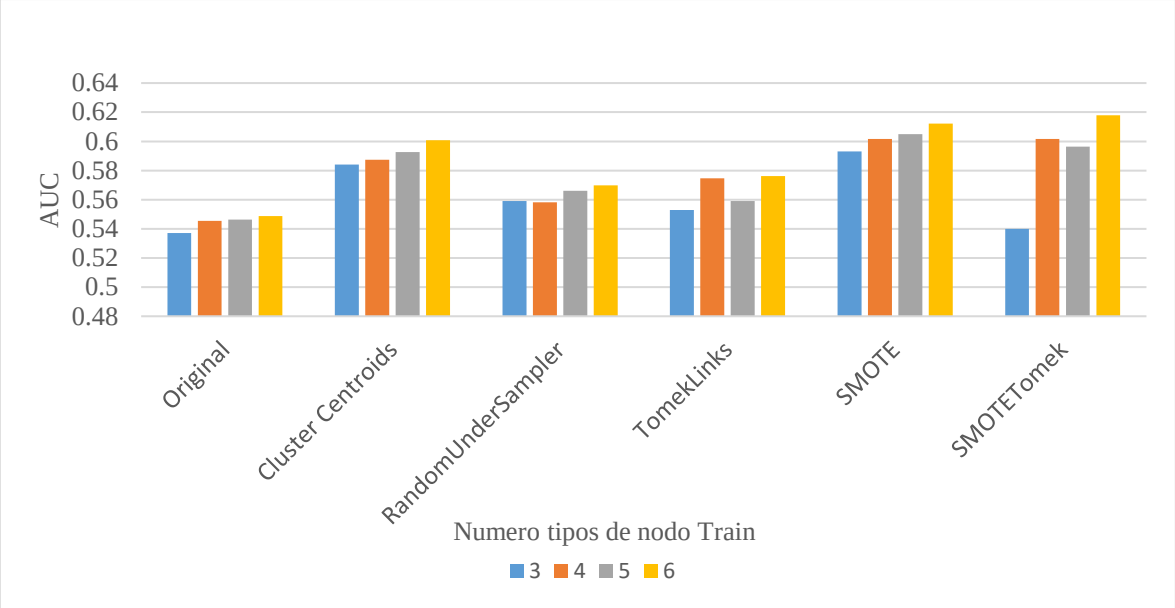
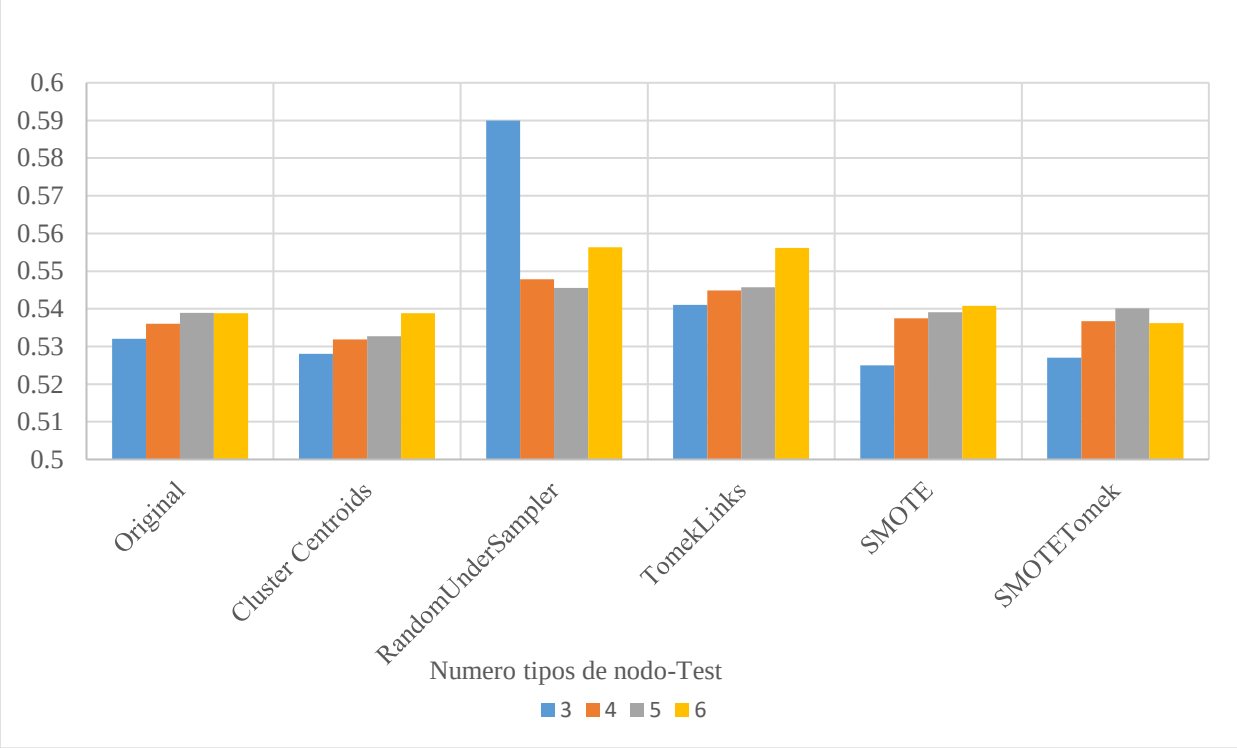


Figura 4
DESEMPEÑO POR NUMERO DE TIPOS DE NODOS (TEST)



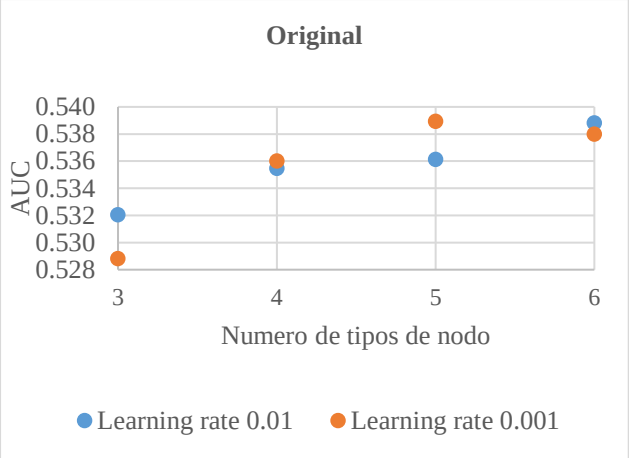
4.5. Learning rate

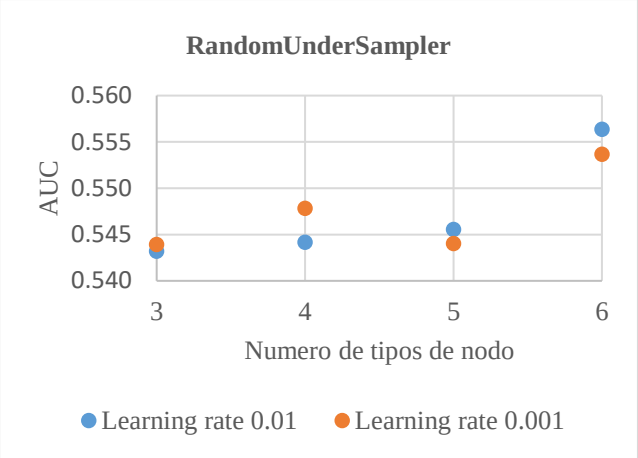
En el tuning de los modelos se encontró dos parámetros que afectaban fuertemente los modelos. El primero fue el learning rate que solo obtenía resultados con capacidad de discriminación en dos valores (0.01 y 0.001). Las pruebas con este parámetro generaban que un método tuviera resultados

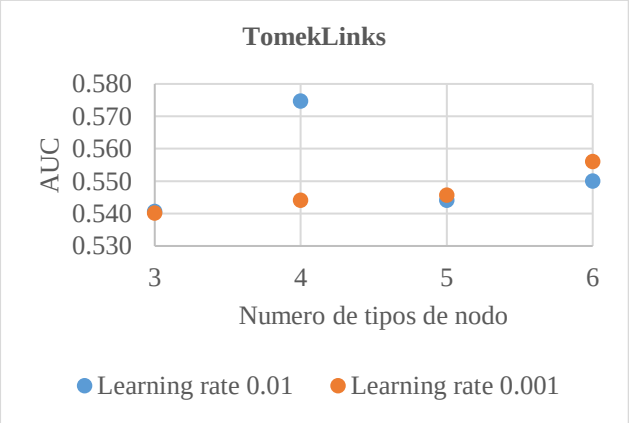
significativos o resultados sin capacidad de discriminación. Es interesante destacar que en promedio la tendencia a la mejora con forme se añadían nodos se mantuvo, además se logró apreciar que a mayor cantidad de tipos de nodos los modelos predecían en mayor frecuencia la clase positiva. Esto es un factor importante para analizar en el desarrollo de modelos de grafos heterogéneos, ya que si se tiene como métrica de importancia la predicción de la clase positiva se deben buscar formas de enriquecer el grafo, esto es aún más importante teniendo en cuenta que el enriquecimiento del grafo mediante la adhesión de tipos de nodos se puede hacer reconvirtiendo características internas de algún tipo de nodo existente. Esto permite una especie de data augmentation sin la necesidad de recopilar nuevos datos, siendo así los grafos una estructura con capacidad de reestructuración para la mejora del desempeño.

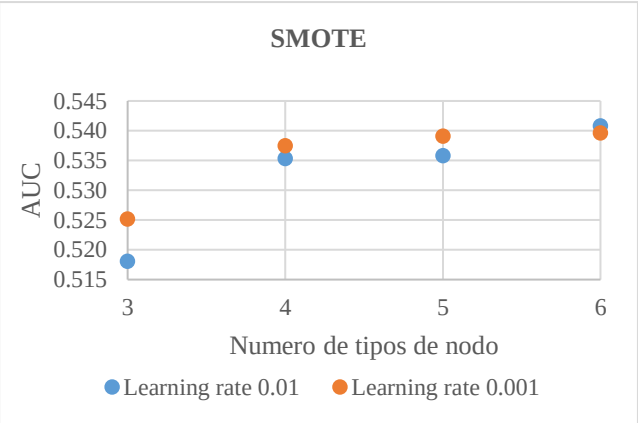
Figura 5

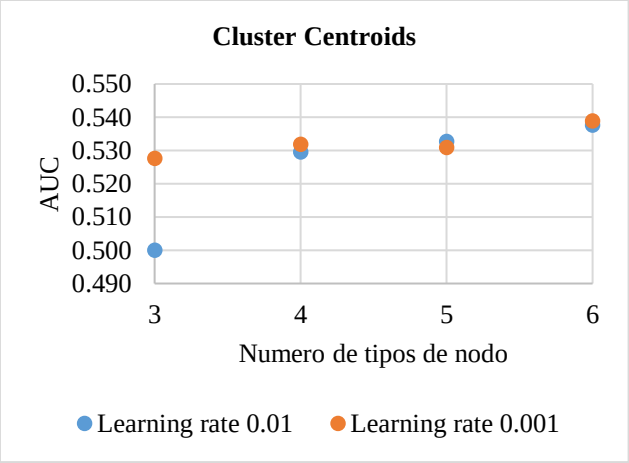
COMPARACION LEARNING RATE

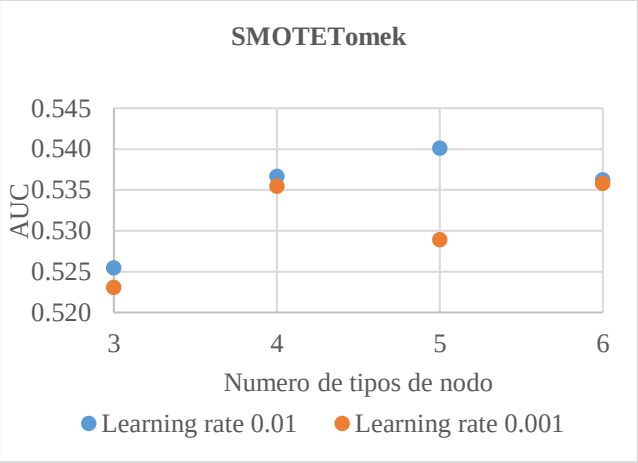












4.6 Epoch

El segundo parámetro de interés fue el número de epochs. Este fue importante no por su mejora en el AUC si no por la modificación que genera en el recall, se pudo observar que el epoch con

mejor AUC no correspondía con el epoch con mejor recall. La modulación de los epoch permitió encontrar un parámetro de ajuste en el equilibrio entre métricas, ya que AUC elevados que no corresponden a discriminación entre clases importante no es un resultado favorable. Es importante destacar que está diferencial generada por el número de epochs es de mayor presencia al momento de la ejecución sobre el conjunto de prueba. Es relevante destacar que el número de epochs no género una diferenciación importante en el desarrollo del overfitting, ya que este se sostenía en un diferencial constante en los número de epochs dónde se encontró

Tabla 3
COMPARACION CANTIDAD DE EPOCH

Resampling method	**Epoch 100**		**Epoch 200**		**Epoch 300**	
	Test		**Test**		**Test**	
	AUC	**Recall**	**AUC**	**Recall**	**AUC**	**Recall**
Original	0.531	0.207	0.532	0.212	0.531	0.208
Cluster Centroids	0.509	0.690	0.524	0.646	0.528	0.598
RandomUnderSampler	0.518	0.586	0.520	0.589	0.590	0.590
TomekLinks	0.531	0.371	0.539	0.410	0.540	0.340
SMOTE	0.502	0.295	0.521	0.348	0.525	0.338
SMOTETomek	0.524	0.445	0.524	0.439	0.527	0.445

4.7 Mejores performance

Teniendo en cuenta todo lo anterior podemos observar que los dos métodos con mejores resultados son las técnicas de undersampling, las cuales obtuvieron un AUC comparable con la línea base y equilibrados en términos de la discriminación entre clases, mostrando además un buen desempeño en la clase positiva.

Tabla 4
RESUMEN MEJORES RESULTADOS

Resampling	**Train**		**Test**	
	AUC	**Recall**	**AUC**	**Recall**
Original	0.537	0.222	0.532	0.212
Cluster Centroids	**0.584**	**0.609**	0.528	0.598
RandomUnderSampler	0.559	0.604	**0.590**	**0.590**
TomekLinks	0.553	0.362	0.541	0.345
SMOTE	0.593	0.479	0.525	0.338
SMOTETomek	0.540	0.457	0.527	0.445

5. CONCLUSIONES

En resumen, este estudio demostró que la utilización de redes neuronales de grafos heterogéneas es prometedora para el análisis de datos médicos estructurados y la predicción de readmisión hospitalaria en pacientes con diabetes. La conversión de la base de datos tabular a una representación de grafo permitió capturar las relaciones complejas entre las variables médicas, lo que condujo a una nueva forma de analizar el problema.

Se observó que la complejidad del grafo, en términos de la adición de diferentes tipos de nodos, tuvo un impacto en el rendimiento del modelo, especialmente en la capacidad de discriminación entre clases. Además, se identificó la importancia de ajustar parámetros clave, como el learning rate y el número de epochs, para lograr un equilibrio entre diferentes métricas de evaluación.

Si bien no se logró cumplir la hipótesis propuesta es de rescatar el performance del random undersampling, el cual es especialmente interesante porque demuestra que un balanceo simple puede generar un gran efecto en la red de grafos, generando un mayor interés en la implementación de esta metodología en una base de datos sin desbalance, con el fin de probaar la hipótesis en una configuración estandar.

En conclusión, este estudio proporciona una base sólida para futuras investigaciones en el campo de las redes neuronales de grafos heterogéneas aplicadas al análisis de datos médicos estructurados. Se recomienda explorar aún más la incorporación de diferentes tipos de nodos y relaciones en los grafos, así como investigar enfoques más avanzados de resampling para abordar el desbalance de clases. Estos avances podrían conducir a mejores modelos de HetGNN y contribuir al desarrollo de herramientas más efectivas para la toma de decisiones en el campo de la medicina.

REFERENCIAS

[1] A. Becker, “Artificial intelligence in medicine: What is it doing for us today?,” Health Policy Technol, vol. 8, no. 2, pp. 198–205, Jun. 2019, doi: 10.1016/J.HLPT.2019.03.004.

[2] M. Singh and K. Kaur, “SQL2Neo: Moving health-care data from relational to graph databases,” Souvenir of the 2015 IEEE International Advance Computing Conference, IACC 2015, pp. 721–725, Jul. 2015, doi: 10.1109/IADCC.2015.7154801.

[3] Wang, X., et al. (2019). Heterogeneous Graph Attention Network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

[4] Hu, M., et al. (2020). Heterogeneous Graph Transformer. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.

[5] Li, Y., et al. (2020). Heterogeneous Graph Attention Network with Gated Recurrent Units. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management.

[6] Zhang, C., et al. (2020). HetGNN: Heterogeneous Graph Neural Network. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

[7] Zhang, P., et al. (2020). Multi-modal Graph Convolutional Networks for Heterogeneous Graphs. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

[8] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” IEEE Trans Neural Netw, vol. 20, no. 1, pp. 61–80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.

[9] C. Zhang, D. Song, C. Huang, A. Swami, and N. v. Chawla, “Heterogeneous graph neural network,” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 793–803, Jul. 2019, doi: 10.1145/3292500.3330961.

[10] R. Ragesh, S. Sellamanickam, A. Iyer, R. Bairi, and V. Lingam, “HeteGCN: Heterogeneous Graph Convolutional Networks for Text Classification”.

[11] Mehrabi, S., Krishnan, A., Hessel, A., & Panchal, J. (2020). Intelligent therapeutic decision support for 30 days readmission of diabetic patients with different comorbidities. IEEE Journal of Biomedical and Health Informatics, 24(11), 3200-3209.

[12] P. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10302

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785-794. [Online].

[14] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. [Online]

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015. [Online]

[16] M. Wang, Y. Cui, J. Zhu, Q. Zhang, and J. Tang, "Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 2019, pp. 1666-1674. doi: 10.1145/3292500.3330974.