Ricardo Mosso Guerrero

# Reporte Final: Análisis de Enfermedades Crónicas No Transmisibles en Guerrero mediante Aprendizaje Profundo

## Justificación del Proyecto

El presente proyecto surge de la necesidad de comprender mejor los patrones y relaciones entre las Enfermedades Crónicas No Transmisibles (ECNT) en el estado de Guerrero, México. Las ECNT, particularmente la diabetes, hipertensión y obesidad, representan un grave problema de salud pública en México, con altas tasas de morbilidad y mortalidad.

La elección de implementar un Autoencoder Variacional (VAE) como metodología principal se fundamenta en varias razones:

1. Complejidad de los datos: El conjunto de datos analizado (geria2023gro.csv) contiene 10,767 registros con 85 variables, de las cuales seleccionamos 46 relacionadas específicamente con diabetes (22), hipertensión (12) y obesidad (12). Esta alta dimensionalidad dificulta el análisis tradicional y requiere técnicas avanzadas de reducción de dimensionalidad.
2. Alto porcentaje de valores faltantes: Los datos presentan un promedio de 56% de valores faltantes por columna, con valores que oscilan entre 4.66% y 99.06%. Los VAE son particularmente robustos frente a datos incompletos, ya que pueden aprender representaciones latentes significativas incluso con información parcial.
3. Necesidad de identificar patrones ocultos: Las comorbilidades entre ECNT siguen patrones complejos y no lineales que no son fácilmente detectables mediante técnicas estadísticas convencionales. El VAE permite descubrir estas relaciones subyacentes al proyectar los datos en un espacio latente de menor dimensión.
4. Potencial para la segmentación de pacientes: El espacio latente generado por el VAE facilita la identificación de grupos naturales de pacientes con perfiles similares de ECNT, lo que puede informar intervenciones de salud pública más específicas y personalizadas.

## Metodología Implementada

Implementamos un Autoencoder Variacional (VAE) utilizando tecnologías de aprendizaje profundo. El modelo se entrenó durante 50 épocas, monitoreando tanto la pérdida de entrenamiento como la de prueba. La arquitectura del VAE incluyó:

1. Un codificador que comprime los datos de entrada de 46 dimensiones a un espacio latente bidimensional.
2. Un decodificador que reconstruye los datos originales a partir de la representación latente.
3. Una capa de muestreo estocástico que garantiza la continuidad del espacio latente.

Complementamos el análisis del VAE con técnicas de agrupamiento (clustering) aplicadas al espacio latente para identificar patrones naturales en los datos.

Resultados Obtenidos

Rendimiento del Modelo

El entrenamiento del VAE mostró una convergencia satisfactoria, con una reducción constante en la función de pérdida tanto para el conjunto de entrenamiento como para el de prueba. La pérdida de entrenamiento disminuyó de aproximadamente 40 en la primera época a 24.82 en la época 50, mientras que la pérdida de prueba se redujo de 35 a 29.01.

La diferencia persistente entre la pérdida de entrenamiento y la de prueba (aproximadamente 4.19 unidades en la época 50) indica cierto grado de sobreajuste, aunque la tendencia decreciente en ambas curvas sugiere que el modelo sigue aprendiendo patrones generalizables.

Visualización del Espacio Latente

La proyección de los datos en el espacio latente bidimensional reveló estructuras interesantes:

1. La mayoría de los puntos se concentran en una región central, formando un núcleo denso.
2. Existen varios puntos atípicos (outliers) dispersos en el espacio, particularmente en las regiones con valores $z[0] > 10$ y $z[1] < -10$.
3. No se observa una separación clara entre los conjuntos de entrenamiento y prueba, lo que sugiere que el modelo ha capturado patrones consistentes en ambos conjuntos.

Identificación de Clusters

Ricardo Mosso Guerrero

El análisis de agrupamiento aplicado al espacio latente identificó tres clusters principales:

1. Un grupo mayoritario (en turquesa) concentrado en la región central superior.
2. Un segundo grupo (en púrpura) ubicado en la región central inferior.
3. Algunos puntos aislados que podrían representar casos atípicos o subgrupos minoritarios.

Las características más distintivas entre estos clusters están relacionadas principalmente con las variables AOB03, ADM03, AHA03, AHA02 y AOB05, lo que sugiere que ciertos indicadores específicos de las tres ECNT estudiadas son particularmente relevantes para la segmentación de los pacientes.

Conclusiones

A partir del análisis realizado, podemos extraer las siguientes conclusiones:

1. Patrones de comorbilidad identificados: El VAE ha logrado capturar estructuras significativas en los datos que reflejan patrones de comorbilidad entre diabetes, hipertensión y obesidad. La distribución de los puntos en el espacio latente y la formación de clusters sugieren la existencia de perfiles distintivos de pacientes.
2. Importancia de variables específicas: Las variables con el sufijo "03" en las tres categorías de ECNT (ADM03, AHA03, AOB03) emergen consistentemente como las más importantes para distinguir entre grupos de pacientes. Esto sugiere que ciertos indicadores específicos tienen un valor predictivo superior y deberían recibir atención prioritaria en la práctica clínica.
3. Desafíos en la calidad de los datos: El alto porcentaje de valores faltantes (56% en promedio) representa una limitación significativa para el análisis. A pesar de esto, el VAE ha demostrado robustez al capturar patrones coherentes, aunque la interpretación de los resultados debe hacerse con cautela.
4. Convergencia del modelo: La disminución constante en las curvas de pérdida indica que el modelo ha logrado aprender representaciones útiles de los datos. Sin embargo, la brecha persistente entre las pérdidas de entrenamiento y prueba sugiere que hay margen para mejorar la generalización del modelo.
5. Potencial para la estratificación de riesgos: Los clusters identificados en el espacio latente ofrecen una base prometedora para desarrollar sistemas de estratificación de riesgos que podrían informar intervenciones preventivas y terapéuticas más precisas.

Ricardo Mosso Guerrero

Recomendaciones y Trabajo Futuro

1. Mejorar la recolección de datos: Es fundamental reducir el porcentaje de valores faltantes en futuros registros para obtener resultados más confiables y representativos.
2. Validación clínica: Los patrones identificados por el VAE deberían validarse mediante la evaluación clínica para confirmar su relevancia médica y potencial aplicación en la práctica.
3. Expansión del modelo: Incorporar variables socioeconómicas, demográficas y de estilo de vida podría enriquecer el análisis y proporcionar una comprensión más holística de los factores que influyen en las ECNT.
4. Implementación de un sistema predictivo: Desarrollar un sistema que utilice el espacio latente del VAE para predecir la progresión de las ECNT y el riesgo de complicaciones.
5. Análisis longitudinal: Extender el estudio para incluir datos temporales permitiría comprender mejor la evolución de las ECNT y las trayectorias de los pacientes a lo largo del tiempo.

En conclusión, este proyecto ha demostrado el valor de las técnicas avanzadas de aprendizaje profundo, específicamente los Autoencoders Variacionales, para el análisis de datos complejos en el ámbito de la salud pública. A pesar de las limitaciones en la calidad de los datos, hemos logrado identificar patrones significativos que pueden informar estrategias más efectivas para abordar la creciente carga de las Enfermedades Crónicas No Transmisibles en Guerrero.