Aplicación de algoritmo Isolation Forest para detección de fraude en reclamos médicos

Por: Mateo Rodríguez Salguero

Realidad del Ecuador

- Según la ACFE se pierde alrededor del 5% de los ingresos anuales de una organización por fraudes.
- En el Ecuador o LATAM no existen cifras oficiales que estimen las pérdidas por fraude en las compañías.
- Existen nuevos esquemas de fraude debido a la implementación de tecnologías (incremento por pandemia).

Cresswell R. (2018). Health care fraud: 5 common billing schemes. ACFE Insights. https://www.acfeinsights.com/acfe-insights/2018/12/12/health-care-fraud-5-common-billing-schemes

Enfoques para detección de fraude en el sector

Debido a la transformación digital que existe en las empresas es necesario emplear técnicas que permitan el monitoreo continuo de procesos.

Comparación entre enfoques

Uso de herramientas tecnológicas para el análisis de datos y técnicas de machine learning para predicciones

Tradicional

Estudios actuariales basados en riesgos

Análisis descriptivos

Revisión documental

Canal de denuncias

Actual

Implementación de herramientas de inteligencia artificial

Modelos predictivos (ML)

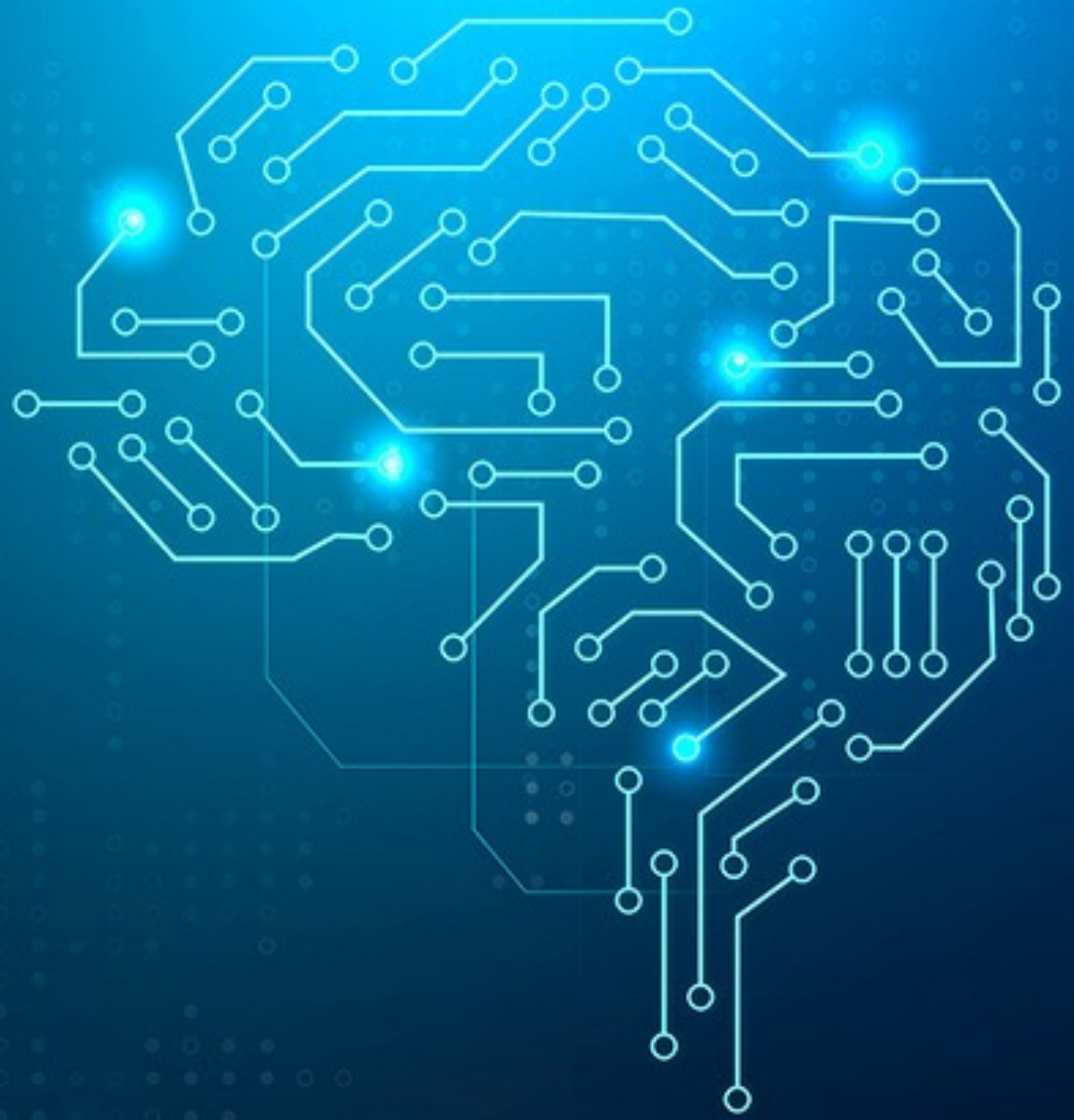
Análisis de patrones

Procesamientos de lenguaje natural

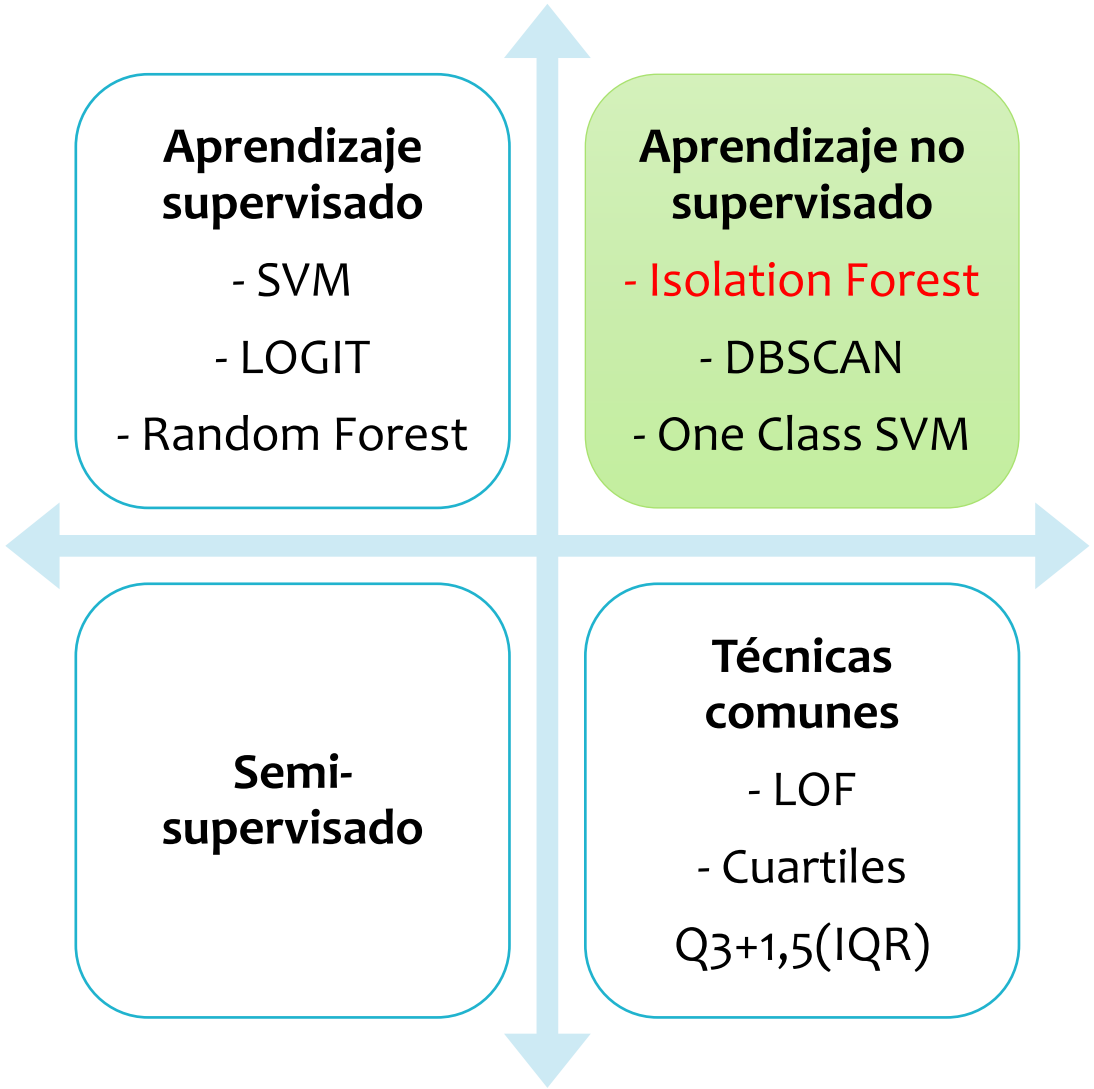
MONITOREO CONTINUO

Métodos a utilizar para detección de anomalías (fraude)

Aprendizaje supervisado, no supervisado, semi-supervisado, métodos comunes



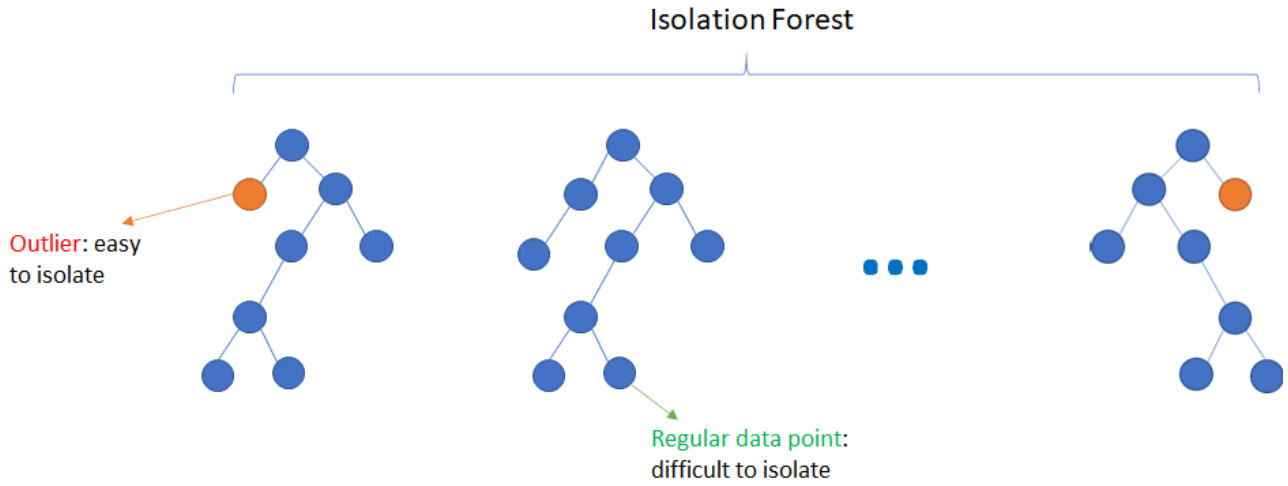
ALGORITMOS PARA LA DETECCIÓN DE FRAUDE



Isolation Forest para detección de anomalías (outliers)

Funcionamiento Isolation Forest:

- Método de detección de anomalías (outliers) no supervisado.
- Estructura basada en árboles lo que permite aislar los registros considerados como anómalos.
- Calcula un valor aleatorio dentro del rango de una variable (Ese valor será el valor de partición para el árbol).
- Se adapta a grandes conjuntos de datos



Numero de árboles – Contaminación - Bootstrap

Vijayakumar, V., Nallam Sri Divya, Sarojini, P. & Sonika, K. (2020). Isolation Forest and Local Outlier Factor for credit card fraud detection system. International Journal of Engineering and Advanced Technology, 9(4), 261–265. https://doi.org/10.35940/ijeat.d6815.049420

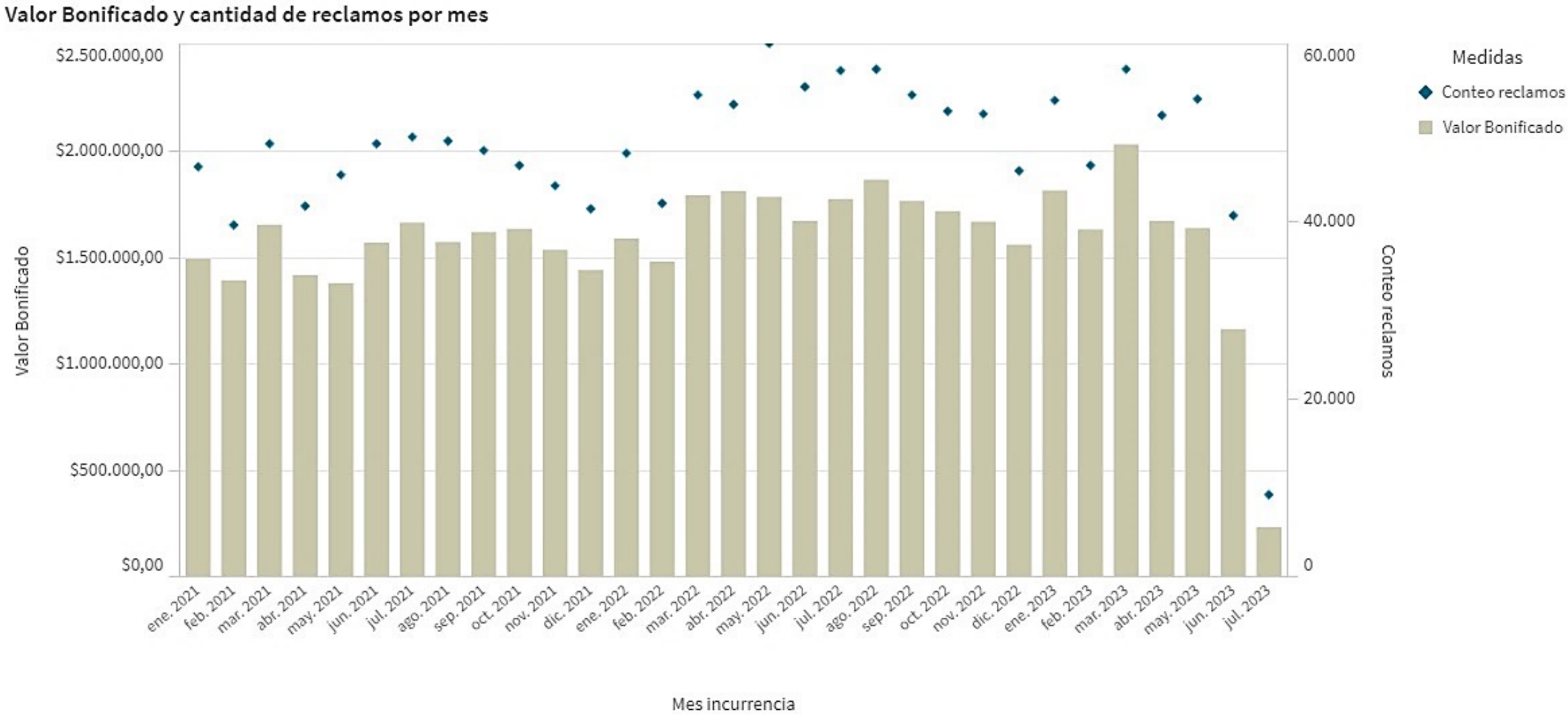
Aplicación en una empresa de medicina prepagada

EDA – Modelado - Resultados

Análisis de los reembolsos presentados a la compañía

Datos tomados de la organización por reclamos pagados de contado a clientes que presentaron atenciones médicas desde el 1 de enero del 2021 a julio del 2023

Análisis de los reembolsos presentados a la compañía



Datos tomados de la organización por reclamos pagados de contado a clientes que presentaron atenciones médicas desde el 1 de enero del 2021 a julio del 2023

Selección de variables para el modelo

Variables cuantitativas

- Valor presentado por la atención medica
- Cantidad de procedimientos que se realiza
- Monto de cobertura del plan médico
- Periodo en días hasta la presentación del reclamo
- Edad del beneficiario

Variables Cualitativas

- Grupo del diagnóstico médico
- Nivel del prestador médico (Minimo-máximo)
- Lugar de atención
- Código del producto o plan de medicina
- Región de donde proviene el contrato
- Diagnóstico médico principal
- Beneficio cubierto por la aseguradora
- Tipo de prestador

EDA

- Corrección de registros
- Unificación de valores
- Eliminar datos incoherentes

Preparación de datos

- Estandarización de textos
- Codificar variables
- Escalar variables

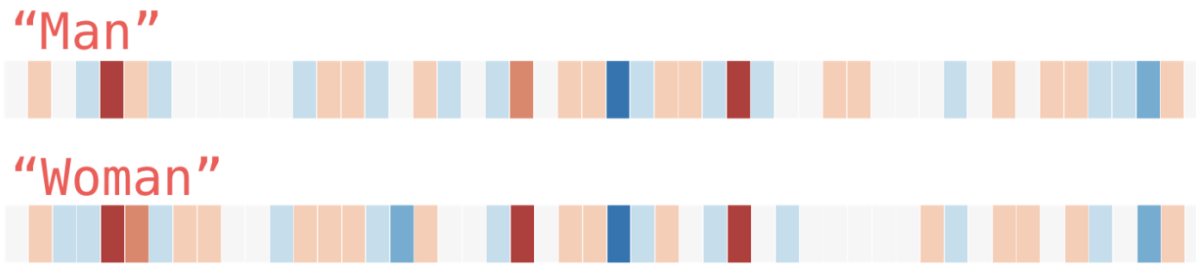
Modelado de datos

- Isolation Forest
- Evaluación de scores de anomalías
- Análisis de resultados

Word Embeddings para aportar información al modelo

Word Embeddings aplicando word2vec:

Este término hace referencia a vectorizar las palabras para ubicarlas dentro de un espacio y que puedan aportar información.



WORD 2 VEC

Skip-gram

Jay was hit by a _____ bus in...

by	a	red	bus	in

CBOW (Continous –bag-of-words)

Jay was hit by a red bus in...

by	a	red	bus	in

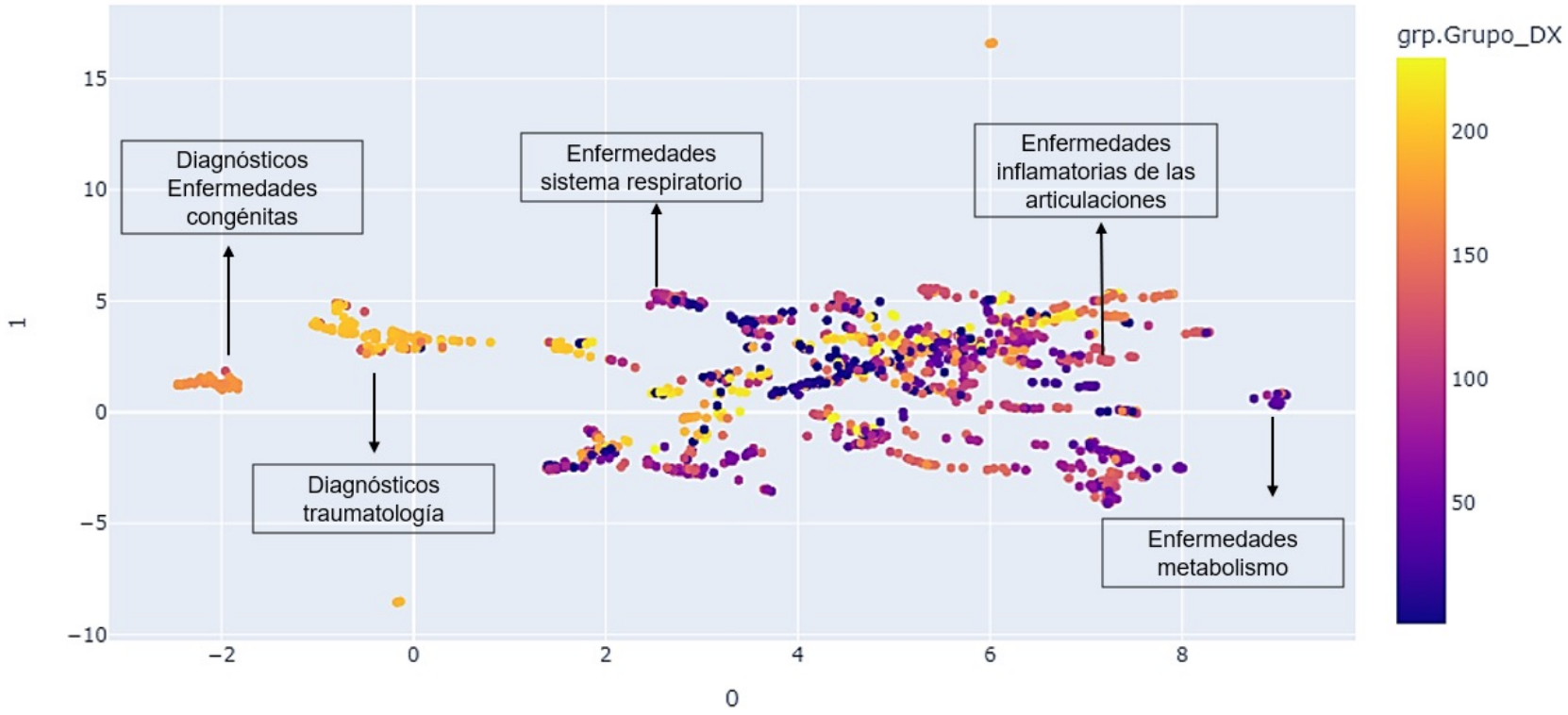
input	output
red	by
red	a
red	bus
red	in

1) Naili, M., Chaibi, A. H., & Ben Ghezala, H. H. (2017). Comparative study of word embedding methods in topic segmentation. Procedia Computer Science, 112, 340–349. https://doi.org/10.1016/j.procs.2017.08.009

2) Alammar, J. (s/f). The Illustrated Word2vec. Github.io. Recuperado el 23 de agosto de 2023, de https://jalammar.github.io/illustrated-word2vec/

Vector de diagnóstico médico reducido a 2 dimensiones

Diagnóstico	**Vector de tamaño 15 (DX1+DX2+Dx3)**
Rinofaringitis	[-1.9758183 -1.0550183 -0.44513655 1.5 66888 …]
Aguda	[-1.6970391 1.2559682 -0.30808872 0.35042754 …]
(Rinofaringitis + aguda)/2	[-1.8364286 0.10047495 -0.37661263 0.95865774 …]



Evaluación de resultados

ISOFOR sin variable de diagnostico

4 Casos de fraude comprobado por procedimientos médicos no realizados

46 Casos de alerta de un prestador por posibles procedimientos médicos sin cobertura o upcoding.

ISOFOR variable de diagnóstico codificada

8 Casos de fraude comprobado por procedimientos médicos no realizados

133 Casos de alerta de un prestador por posibles procedimientos médicos sin cobertura o upcoding.

Resultados con la variable diagnóstico vectorizada

LR02.CruceLR02	PR07.nombre-beneficio	LR04.valor-presentado	LR04.cantidad-presentada	MONTO DE COBERTURA	LR02.lugar-atencion	periodo_presentacion	LR13.Nombre Cabecera	LR02.Edad beneficiario	LR02.codigo-producto	LR02.region	grp.Grupo_DX	CO03.tipo-prestador	CO03.nivel-prestador-desde	CO03.nivel-prestador-hasta	diagnostico_codificado	anomaly	score
27996802-0	HONORARIOS MEDICOS	2520.00	12.0	30000	CONSULTA EXTERNA	75	VERRUGAS VIRICAS	22	IND	COSTA	12.0	MEDICO	6.0	6.0	[2.026625633239746, -0.09704160690307617, 3.33...	-1	-0.006236
28109702-0	HONORARIOS MEDICOS	3875.00	5.0	30000	CONSULTA EXTERNA	89	VERRUGAS VIRICAS	33	IND	COSTA	12.0	MEDICO	6.0	6.0	[2.026625633239746, -0.09704160690307617, 3.33...	-1	-0.001068
28113552-0	HONORARIOS MEDICOS	200.00	20.0	1000000	CONSULTA EXTERNA	27	BRONQUITIS AGUDA	34	XPR	COSTA	98.0	MEDICO	6.0	6.0	[-4.674031138420105, 2.7726617455482483, 0.962...	-1	-0.000030
598199111-2	HONORARIOS MEDICOS	1425.00	75.0	45000	CONSULTA EXTERNA	90	VERRUGAS VIRICAS	3	IND	COSTA	12.0	MEDICO	6.0	6.0	[2.026625633239746, -0.09704160690307617, 3.33...	-1	-0.021780
598411404-0	HONORARIOS MEDICOS	2520.00	150.0	15000	CONSULTA EXTERNA	55	VERRUGAS VIRICAS	53	IND	COSTA	12.0	MEDICO	6.0	6.0	[2.026625633239746, -0.09704160690307617, 3.33...	-1	-0.012812
598458271-0	HONORARIOS MEDICOS	2520.00	12.0	30000	CONSULTA EXTERNA	74	VERRUGAS VIRICAS	42	IND	COSTA	12.0	MEDICO	6.0	6.0	[2.026625633239746, -0.09704160690307617, 3.33...	-1	-0.007112
598610240-2	HONORARIOS MEDICOS	514.52	1.0	15000	HOSPITAL	31	POLIPO DEL TRACTO GENITAL FEMENINO	43	IND	COSTA	143.0	MEDICO	6.0	6.0	[-2.7772019505500793, -3.4414150118827820, -1.9...	-1	-0.007397
598744102-0	HONORARIOS MEDICOS	500.00	2.0	30000	HOSPITAL	27	TRASTORNOS NO INFLAMATORIOS DEL OVARIO, DE LA ...	32	IND	COSTA	143.0	MEDICO	6.0	6.0	[1.1454779846327645, -5.562457391193935, -3.33...	-1	-0.001971

Conclusiones

- Utilizar algoritmos de machine learning puede mejorar la gestión de las áreas de auditoría y control (generar ahorro).
- ISOFOR realiza correctamente su trabajo de aislamiento para detectar casos que no sean comunes en las prestaciones médicas y puedan ser fraudes.
- Existe un mejor rendimiento del modelo al agregar la variable de diagnóstico vectorizada.
- El uso de variables cualitativas en detección de fraude ayuda a segmentar de una mejor manera los diferentes casos y tener una mayor precesión del modelo.

Recomendaciones

- Añadir variables como especialidad del doctor y en el caso que aplique si la atención se da en una clínica de especialidad.
- Evaluar la implementación de variables que indiquen un riesgo en la atención basado en características de prestadores o clientes con irregularidades.
- Trabajar con la creación de una base de casos de fraude para trabajar con algún modelo supervisado o semi-supervisado.

Referencias:

- Carletti, M., Terzi, M., & Susto, G. A. (2023). Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest. Engineering Applications of Artificial Intelligence, 119(105730), 105730. https://doi.org/10.1016/j.engappai.2022.105730
- Cresswell R. (2018). *Health care fraud: 5 common billing schemes*. ACFE Insights. https://www.acfeinsights.com/acfe-insights/2018/12/12/health-care-fraud-5-common-billing-schemes
- Churgin, M & Bansal, J. (2022, julio 19). Embedding medical journeys with machine learning to improve member health at CVS Health. CVS Health Tech Blog. https://medium.com/cvs-health-tech-blog/embedding-medical-journeys-with-machine-learning-to-improve-member-health-at-cvs-health-957148339cd6
- Mehbodniya, A., Alam, I., Pande, S., Neware, R., Rane, K. P., Shabaz, M., & Madhavan, M. V. (2021). Financial fraud detection in healthcare using machine learning and deep learning techniques. *Security and Communication Networks*, *2021*, 1–8. https://doi.org/10.1155/2021/9293877
- Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. International Journal of Information Management Data Insights, 1(2), 100012. https://doi.org/10.1016/j.jjimei.2021.100012
- Rukhsar, L., Haider Bangyal, W., Nisar, K., Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University research journal of engineering and technology, 41(1), 33–40. https://doi.org/10.22581/muet1982.2201.04
- Settipalli, L., & Gangadharan, G. R. (2023). WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims. Expert Systems with Applications, 215(119259), 119259. https://doi.org/10.1016/j.eswa.2022.119259

Gracias

Mateo Rodríguez
0998544853
mateo.rodriguez.salguero@udla.edu.ec