Ensayo de Investigación

Comparación de algoritmos de machine learning en la clasificación de la hipertensión

Comparison of machine learning algorithms in hypertension classification

David Mauricio Cañedo Figueroa[1], Carlos Eduardo Cañedo Figueroa[2]*

[1]Facultad de Medicina
Universidad Autónoma de Sinaloa
[2]Facultad de Medicina y Ciencias Biomédicas
Universidad Autonoma de Chiahuahua

Autor de correspondencia:
*ccanedo@uach.mx

Recibido: 08-01-2024 Aceptado: 07-04-2025 (Artículo Arbitrado)

Resumen

En el presente documento se redacta la implementación de 10 algoritmos de aprendizaje automático (ML) con el objetivo de clasificar datos clínicos y de laboratorios asociados a el riesgo hipertensión en pacientes mexicanos, datos disponibles en el sitio web Kaggle con la búsqueda "hipertensión arterial México Data Set", entre los principales están Decision tree (DT) (99 %), Random Forest (RM) (98 %), red neuronal artificial (AAN) (65 %), Naive Bayes (NB) (55 %), K-Nearest Neighbors (KNN) (41 %) y Support Vector Machines (SVM) (45 %) además de generar dos algoritmos ensamblados usando los tres algoritmos con mejor resultados (DT, RM y ANN), el primero con un enfoque de votación con precisión de 98.47 % y el segundo usando stkacking con 99.39 %. Se concluyo que los resultados del análisis experimental muestran la efectividad de cada algoritmo de ML, así como la precisión de los algoritmos ensamblados, proponiendo un uso posterior para centros médicos y epidemiológicos con la utilidad de analizar los factores más importantes en el riesgo de hipertensión.

Palabras clave: Análisis, random forest, decisión tree, red neuronal artificial, mexicanos.

Abstract

Ten machine learning (ML) algorithms were implemented with the aim of classifying clinical and laboratory data associated with hypertension risk in Mexican patients, data available on the Kaggle website with the search "hypertension Mexico Data Set", among the main ones are Decision tree (DT) (99%), Random Forest (RM) (98%), Artificial Neural Network (AAN) (65%), Naive Bayes (NB) (55%), K-Nearest Neighbors (KNN) (41%) and Support Vector Machines (SVM) (45%) in addition to generating two ensemble algorithms using the three best performing algorithms (DT, RM and ANN), the first with a voting approach with accuracy of 98. 47% and the second using stkacking with 99.39%. It was concluded that the results of the experimental analysis show the effectiveness of each ML algorithm, as well as the accuracy of the assembled algorithms, proposing a further use for medical and epidemiological centers with the utility of analyzing the most important factors in the risk of hypertension.

Keywords: Analysis, random forest, decision tree, artificial neuronal network, mexicanos .

Introducción

La hipertensión es una de las anomalías cardiovasculares que más prevalece en diferentes sectores de la población que involucran rangos de edad, raza, nivel socioeconómico y algunas variables como hábitos y costumbres (Martínez-Álvarez, 2023). Es catalogada como una enfermedad crónico degenerativa, ya que conlleva una degradación física y mental que provoca un desequilibrio con afectación directa e indirectamente en algunos órganos y tejidos en quien la padece (Guamán Tacuri & López Pérez, 2023). Se puede describir como una condición clínica en donde, el paciente presenta alteración de la presión arterial elevándola de manera crónica, esta definición de elevación se relaciona principalmente con los valores de presión sistólica y presión diastólica. Se considera que un paciente presenta una presión alta, cuando se presentan valores iguales o superiores a 130 en sistólica y 80 en diastólica, ambas medidas en mmHg (Chaulin et al., 2021).

Se ha establecido que, aunque la lectura de la presión es un indicativo razonable para intuir si se sufre o no de hipertensión, se ha reportado que es necesario evaluar una serie de variables tanto clínicas como de laboratorio que confirmen este diagnóstico, factores tales como el tipo de dieta, la cantidad de ejercicio realizado y estudios de biomarcadores como perfil de lípidos o análisis hormonales (N. Nasir et al., 2021).

Según la encuesta nacional de Salud y nutrición (ENSANUT) en el año 2024, la prevalencia de hipertensión arterial fue de 29.9 %, de los cuales el 27.5 % fue de mujeres y el 32.5 % fue de hombres. Se resalta que un 43 % padecían hipertensión no diagnosticada y el 36.3 % de los pacientes diagnósticados se encontraban bajo tratamiento (Campos-Nonato et al., 2023). Esto resalta una necesidad imperiosa de generar diagnósticos más eficientes y con variables diferentes que coayuden al personal de salud.

Se ha mencionado que el diagnóstico pude llegar a ser mutable, pues se presentan nuevos estudios y metodologías para mejorar su precisión, en este contexto, metodologías basadas en la aplicación de algoritmos permiten determinar la hipertensión en zonas pulmonares mediante sistemas de imagenología como la resonancia magnética o la ecografía usando datos pocos utilizados para el diagnóstico de hipertensión (Campos-Nonato et al., 2024; Lee & Park, 2015).

Algunos algoritmos como las máquinas de soporte vectorial (SVM), redes neuronales convolucionales, algoritmos de K vecinos más cercanos (KNN), algoritmos bayesianos y algoritmos de conjuntos (Ensemble learning) se han utilizado para determinar diferentes estados de salud mendiante el procesamiento de datos, muestras y entrenamientos para poblaciones específicas, demostrando que los algoritmos matemáticos pueden ser aplicados en diferentes contextos médicos (Almonacid et al., 2021; Bhimavarapu et al., 2024; N. Nasir et al., 2021).

En el presente documento se describe el análisis de datos obtenidos de la base de datos Kaggle de población mexicana que indica riesgo o no riesgo de padecer hipertensión, esto empleando algoritmos de machine learnging (ML) tales como Random Forest (RM), Decision Tree (DT), Redes Neuronales Artificial (AAN), Naive Bayes (NB), K-Nearest Neighbors (KNN) y Support Vector Machines (SVM), esto con la intención de poder proveer una primera aplicación de algoritmos ML a datos con tantas variables clínicas diferentes, y presentar un precedente para su posterior aplicación en clínica o epidemiología.

Los algoritmos desarrollados se desarrollaron en el lenguaje de programación Python, utilizando los servidores de Google Colab como el proveedor de los recursos informáticos necesarios para su ejecución.

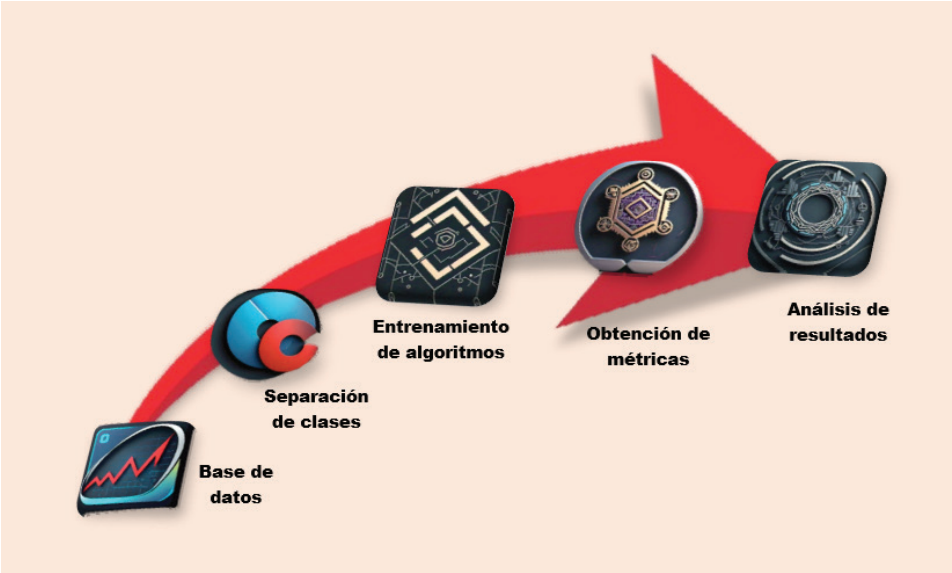

Figura 1. Esquema representativo de la metodología utilizada.
Fuente: Elaboración propia.

Metodología

Los algoritmos desarrollados se escribirán en lenguaje de programación Python, utilizando los servidores de Google Colab como el proveedor de los recursos informáticos necesarios para su ejecución.

La metodología utilizada se puede observar de manera gráfica en la Figura 1, teniendo como primer paso la adquisición de la base de datos, posteriormente se separaron las clases con el fin de preparar el entrenamiento de algoritmos de forma supervisada, posteriormente al entrenamiento, se obtuvieron métricas de rendimiento que evalúan cada algoritmo y finalmente se analizaron los resultados para generar nuevos algoritmos ensamblados. Los algoritmos seleccionados para el desarrollo de este análisis se eligieron de manera puramente experimental.

Adquisición de la base de datos

La base de datos que se utilizó se encuentra alojada en la plataforma Kaggle bajo el nombre de "hipertensión arterial en México"(Felix Jiménez & Sánchez Lee, 2024), dicha base de datos contiene información de 4363 sujetos de prueba identificados con riesgo de hipertensión (2816 datos) o no riesgo de hipertensión (1547). Se realizó la separación del 30 % de los datos (1309 datos, 849 de hipertensión y 460 de no hipertensión) los cuales se utilizaron para probar a los algoritmos una vez que se entrenaron. El 70 % restante de los datos se utilizaron para entrenar a los algoritmos. Para evitar sesgos durante el entrenamiento se balancearon los datos de forma aleatoria dejando un total de 1087 datos de cada una de las clases (hipertensión y no hipertensión).

Es importante mencionar que la base de datos tiene un total de 36 columnas por cada vector característico, las cuales se indican en la Tabla 1. Para el entrenamiento de los algoritmos se descartó la variable "Folio" y se utilizó el entorno de programación de Google Colab mediante el lenguaje de programación de Python.

Tabla 1. Características de la base de datos.

Número de característica	Característica
1	FOLIO_I
2	Sexo
3	Edad
4	Concentración hemoglobina
5	Temperatura ambiente
6	Valor ácido úrico
7	Valor albumina
8	Valor colesterol HDL
9	Valor colesterol LDL
10	Valor colesterol total
11	Valor creatina
12	Resultado glucosa
13	Valor insulina
14	Valor triglicéridos
15	Resultado glucosa promedio
16	Valor hemoglobina glucosilada
17	Valor ferritina
18	Valor folato
19	Valor homocisteína
20	Valor proteína C reactiva
21	Valor transferrina
22	Valor vitamina B12
23	Valor vitamina D
24	peso
25	estatura
26	Medida cintura
27	Segunda medición peso
28	Segunda medición estatura
29	Distancia rodilla - talón
30	Circunferencia de la pantorrilla
31	Segunda medición cintura
32	Tensión arterial
33	Sueno horas
34	Masa corporal
35	Actividad total
36	Riesgo de hipertensión

Fuente: Elaboración propia.

Algoritmos y entrenamiento

Random Forest

El algoritmo Random Forest fue el primero que se evaluó, pues permite clasificar los datos y generar predicciones según las derivaciones que se obtengan, en este trabajo se utilizaron 5 árboles entrenados mediante la función *RandomForestClassifier* de la librería sklearn en Python (Breiman et al., 2003; Celine Vens & Celine Vens, n.d.; Rigatti, 2017).

K Vecinos más cercanos (KNN)

En este trabajo, el algoritmo KNN se entrenó con la función *KNeighborsClassifier* perteneciente a la librería *sklearn*, para esto se utilizó la votación de los 5 vecinos más cercanos al vector utilizado como muestra.

Suport Vector Machine (SVM)

Para la realización del algoritmo SMV, se utilizó un kernel lineal, y este fue entrenado con la función *svm.fit.*

SRedes neuronales artificiales (ANN)

El desarrollo de la red neuronal se decidió por un desarrollo experimental totalmente conectada hacia adelante (feedforward), se desarrollaron 30 neuronas en la capa oculta, con 34 entradas y 2 salidas.

La red neuronal fue entrenada mediante el algoritmo de retro propagación (backpropagation) a 500 épocas. Para ello se utilizó la función *neural_network MLPClassifier* de *sklearn*.

Naive Bayes

En el caso del presente trabajo se implementó el algoritmo bayesiano mediante la función *GaussianNB* que genera un modelo probabilístico que se puede utilizar para clasificación de datos.

Decision tree

En el caso del presente trabajo se utilizó la función DecisionTreeClassifier para la generación del modelo.

Ensamble learning

Existen diversos métodos para generar algoritmos conjuntos, en el presente trabajo se utilizaron dos algoritmos conjuntos, algoritmo de votación (*VotingClassifier*) y algoritmos de stacking (*StackingClassifier*), para ambos algoritmos se utilizaron los algoritmos de mejor métricas obtenidas, los cuales fueron el árbol de decisión, red neuronal y random forest. Para el algoritmo Stacking se utilizó como metamodelo una regresión lineal.

$$Precisión = \frac{TP}{FP + TP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1score = 2 \times \left(\frac{Presición \times Recall}{Presición + Recall} \right) \quad (3)$$

Resultados y discusión

Los algoritmos generados se evaluaron mediante las funciones matemáticas de *Precisión* (1), *Recall* (2) y *F1 score* (3). En donde TP son los vectores clasificados como verdaderos positivos, FN falsos negativos y FP falsos positivos. En la Tabla 2 se muestran los resultados de la evaluación los principales algoritmos utilizados, así como los valores obtenidos de las diferentes métricas evaluadas.

Los resultados de precisión de todos los algoritmos presentan un valor >0.7, sin embargo, son RF, DT y ANN los que presentan un mayor valor de *Precisión*, *Recall* y *F1 score*, sugiriendo que son los mejores modelos para la evaluación de los datos.

Por su parte, los resultados de los conjuntos se presentan en la Tabla 3, en donde, se puede observar que el algoritmo basado en Stacking usando regresión lineal como metamodelo, presentó valores de *Precisión*, *Recall* y *F1 score* >0.98.

Estos resultados indican que el algoritmo en conjunto basado en Stacking tiene un mejor rendimiento al realizar predicciones de la presencia de hipertensión en pacientes con las variables incluidas en este dataset.

Tabla 2. Resultados de la evaluación de algoritmos

Nombre del algoritmo	TP	FP	FN	TN	*Precisión*	*Recall*	*F1 score*
RF	450	10	23	826	0.97	0.975	0.97
KNN	356	104	119	730	0.815	0.815	0.815
SVM	294	166	100	749	0.785	0.76	0.77
ANN	405	55	161	688	0.825	0.845	0.825
Naive Bayes	308	152	114	735	0.78	0.77	0.775
DT	452	8	6	843	0.99	0.985	0.985

Fuente: Elaboración propia.

Tabla 3. Resultados de la evaluación de algoritmo conjunto de dos enfoques

Algoritmo conjunto	TP	FP	FN	TN	*Precisión*	*Recall*	*F1 score*
Voting (RF, ANN, DT)	448	12	14	835	0.98	0.975	0.975
Stacking	452	8	6	843	0.99	0.985	0.985

Fuente: Elaboración propia.

En el año 2023 el equipo de A. Dharma publicó un artículo sobre la identificación de hipertensión utilizando clasificación por Naive Bayes, en donde se presentan valores únicamente de precisión >96 %. Muy similar al trabajo presentado en 2019 por el equipo de Poddar y colaboradores, en donde se realizó un análisis utilizando el algoritmo SVM en la clasificación de hipertensión y arteria coronaria en pacientes, obteniendo resultados similares a los presentados en este artículo con valores de real y precisión ≥ 90 %. Así mismo, en el año 2021 el equipo de Nasair y sus colaboradores, publican una serie de trabajos aplicando algoritmos de ML a pacientes con hipertensión, este trabajo evaluó un conjunto de algoritmos como RF, CatBoost, SVM, KNN, Regresión logística, DT, ANN y XGBoost, siendo solo dos de estos algoritmos con valores de precisión ≥ 90 %. Estos datos se presentan en la Tabla 4.

Conclusiones

Los resultados mostrados en esta investigación muestran un primer acercamiento al uso de ML a dataset complejos y variables para la predicción de hipertensión, muestra que los rendimientos de varios algoritmos evaluados de manera individual (DT, ANN, KNN Y RF) presentan valores de *Precisión*, *Recall* y *F1-score* >0.8 lo que dichita una variedad amplia para la aplicación de modelos de ML a dataset tan heterogéneos. Asimismo, los valores obtenidos de los algoritmos en conjuntos permiten tener una visión más general de cómo se pueden aplicar más de 2 modelos diferentes y evaluar su rendimiento.

Investigaciones a futuro pretenden continuar con evaluaciones más exhaustivas modificando las variables, como por ejemplo evaluando solo biomarcadores, descartando parámetros antropométricos para detectar nuevos biomarcadores asociados al desarrollo de hipertensión los cuales no se tienen registros en la bibliográfica.

Agradecimientos

Se agradece a las universidades autónomas de los estados de Chihuahua y Sinaloa, a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECITHI), de igual manera a las personas involucradas de manera directa e indirecta en la realización de este trabajo.

Referencias

A. Dharma, Anita Sofia Rahmi, Erwin Sitompul, M. Turnip, N. S. Syafei, & A. Turnip. (2023). Hypertension Identification Using Naive Bayes Classification Method and Pan Tompkins Feature Extraction. *International Conference on Telecommunications*. https://doi.org/10.1109/ict60153.2023.10374032

Almonacid, A. B., Rodriguez, C., Pomachagua, Y., & Rodriguez, D. (2021). Hybrid Model based on Support Vector Machine and Principal Component Analysis Applied to Arterial Hypertension Detection. 2021 13th *International Conference on Computational Intelligence and Communication Networks* (CICN), 17–22. https://doi.org/10.1109/CICN51697.2021.9574662

Guamán Tacuri, A. B. & López Pérez G. P. (2023). Hospitalización prevenible en enfermedades crónico degenerativas: Hipertensión arterial y diabetes. *Salud, Ciencia y Tecnología, 3*, 487–487. https://doi.org/10.56294/saludcyt2023487

Bhimavarapu, U., Chintalapudi, N., & Battineni, G. (2024). Automatic Detection and Classification of Hypertensive Retinopathy with Improved Convolution Neural Network and Improved SVM. *Bioengineering, 11*(1), 56. https://doi.org/10.3390/bioengineering11010056

Tabla 4. Resultados publicados por otros autores utilizando modelos de ML a pacientes con hipertensión.

Autor	Algoritmo aplicado	*Precisión*	*Recall*
A. Dharma et al., (2023)	Naive Bayes	96.70 %	N/A
M. G. Poddar et al., (2019)	SVM	96.67 %	90%,
N. Nasir et al., (2021)	Random Forest	90 %	N/A
	CatBoost	87 %	N/A
	SVM	78.33 %	N/A
	KNN	78.33 %	N/A
	Logistic Regression	73.50 %	N/A
	Decision Tree	83.83 %	N/A
	MLP (ANN)	87.33 %	N/A
	XGBoost	90 %	N/A

Fuente: Elaboración propia.

Breiman, L., Last, M., & Rice, J. (2003). Random Forests: Finding Quasars. In *Statistical Challenges in Astronomy* (pp. 243–254). Springer-Verlag. https://doi.org/10.1007/0-387-21529-8_16

Campos-Nonato, I., Oviedo-Solís, C., Hernández-Barrera, L., Márquez-Murillo, M., Gómez-Álvarez, E., Alcocer-Díaz, L., Puente-Barragán, A., Ramírez-Villalobos, D., Basto-Abreu, A., Rojas-Martínez, R., Medina-García, C., López-Ridaura, R., & Barquera, S. (2024). Detección, atención y control de hipertensión arterial. *Salud Pública de México*, *66*(4, jul-ago), 537–546. https://doi.org/10.21149/15867

Campos-Nonato, I., Oviedo-Solís C. I., Vargas-Meza J., Ramírez-Villalobos D., Medi-na-García C., Gómez-Álvarez E., Hernández-Barrera L., & Barquera S. (2023). Prevalencia, tratamiento y control de la hipertensión arterial en adultos mexica-nos: Resultados de la Ensanut 2022. *Salud Pública de México, 65*, s169–s180. https://doi.org/10.21149/14779

Cañedo Figueroa, C. E., & García Chávez, H. (2021). Diseño de algoritmo compuesto por Machine Learning y un modelo probabilístico para la detección de diabetes. *Memorias Del Congreso Nacional de Ingeniería Biomédica, 8*(1), 57–60.

Celine Vens & Celine Vens. (n.d.). Random Forest. *Machine Learning with Regression in Python.* https://doi.org/10.1007/978-1-4419-9863-7_612

Chaulin, A. M., Aleksey M. Chaulin, & Chaulin, A. (2021). Clinical and Diagnostic Value of Highly Sensitive Cardiac Troponins in Arterial Hypertension. *Vascular Health and Risk Management, 17*, 431–443. https://doi.org/10.2147/vhrm.s315376

Dadang Priyanto, Ahmad Robbiul Iman, & Deny Jollyta. (2023). Naïve Bayes and K-Nearest Neighbor Algorithm Approach in Data Mining Classification of Drugs Addictive Diseases. *Ilkom Jurnal Ilmiah, 15*(2), 262–270. https://doi.org/10.33096/ilkom.v15i2.1544.262-270

Felix Jiménez, A. F., & Sánchez Lee, V. S. (2024). *Hipertensión Arterial México* [Data-set]. https://www.kaggle.com/datasets/frederickfelix/hipertensin-arterial-mxico/data

Fennell, B. D., Mezyk, S. P., & McKay, G. (2022). Critical Review of UV-Advanced Reduction Processes for the Treatment of Chemical Contaminants in Water. *ACS Environmental Au, 2*(3), 178–205. https://doi.org/10.1021/acsenvironau.1c00042

Ghassemi, M., Naumann, T., Schulam, P., Chen, I. Y., & Ranganath, R. (n.d.). *A Review of Challenges and Opportunities in Machine Learning for Health.*

Martínez-Álvarez. (2023). Hipertensión. *TEPEXI Boletín Científico de La Escuela Superior Tepeji Del Río, 10*(20), 47–49. https://doi.org/10.29057/estr.v10i20.10818

Lee, J.-H., & Park, J.-H. (2015). Role of echocardiography in clinical hypertension. *Clinical Hypertension*, 21(1), 9. https://doi.org/10.1186/s40885-015-0015-8

M. G. Poddar, Poddar, M. G., Anjali C. Birajdar, Birajdar, A. C., Jitendra Virmani, Virmani, J., . Kriti, & Kriti. (2019). Automated Classification of Hypertension and Coronary Artery Disease Patients by PNN, KNN, and SVM Classifiers Us-ing HRV Analysis. *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*, 99–125. https://doi.org/10.1016/b978-0-12-816086-2.00005-9

Nasir, N., Oswald, P., Barneih, F., Alshaltone, O., AlShabi, M., Bonny, T., & Shammaa, A. A. (2021). Hypertension Classification Using Machine Learning Part II. *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*, 459–463. https://doi.org/10.1109/DeSE54285.2021.9719408

N. Nasir, Omar Alshaltone, Omar Alshaltone, Feras Barneih, Feras Barneih, Mohammad Al-Shabi, Mohammad Al-Shabi, Talal Bonny, Talal Bonny, Ahmed Al-Shammaa, & Ahmed Al-Shamma'a. (2021). Hypertension Classification using Machine Learning—Part I. *International Conference on Developments in eSys-tems Engineering*. https://doi.org/10.1109/dese54285.2021.9719523

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, 47(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

Shahadat Uddin, Shahadat Uddin, Ibtisham Haque, Ibtisham Haque, Haohui Lu, Hao-hui Lu, Mohammad Ali Moni, Mohammad Ali Moni, Ergun Gide, & Ergun Gide. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports, 12*(1). https://doi.org/10.1038/s41598-022-10358-x

Y. Hirata, Takumasa Tsuji, Jun'ichi Kotoku, Masataka Sata, & K. Kusunose. (2024). Echocardiographic artificial intelligence for pulmonary hypertension classification. *Heart*. https://doi.org/10.1136/heartjnl-2023-323320

Yunhua Zhou, Yunhua Zhou, Peiju Liu, Peiju Liu, Xipeng Qiu, & Xipeng Qiu. (2022). KNN-Contrastive Learning for Out-of-Domain Intent Classification. *Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2022.acl-long.352