Predicción de Lenguas en Peligro de Extinción Mediante El Aprendizaje Automático: Una Revisión Sistemática de la Literatura

Predicting Endangered Languages Using Machine Learning: A Systematic Literature Review

Keysi Ariana Ato Fernández

Universidad Nacional Mayor de San Marcos, Lima, Perú

Resumen

Las lenguas en peligro de extinción es un problema que está ocurriendo de forma acelerada debido a diversos factores, según United Nations (2023) más del 43% de lenguas en el mundo se encontraban en peligro de extinción. Identificar la vitalidad de una lengua es una tarea complicada e importante para mantener la diversidad cultural, y debido a que va a depender de muchas causas y la accesibilidad a los datos, puede ser no del todo predecibles. Esta tarea de predecir la vitalidad de un lenguaje puede ser más eficiente y precisa gracias a las técnicas de machine learning. En este sentido, se realizó una revisión sistemática de la literatura en modelos de predicción de la extinción de una lengua basando la búsqueda en cinco preguntas de investigación utilizando la metodología PRISMA. Este estudio científico recuperó 962 artículos de las bases de datos Scopus, IEEE y ScienceDirect. El número de estudios utilizados para responder las preguntas de investigación fue 33.
Finalmente, analizando los Modelos de aprendizaje de máquina, el Generalized Linear Mixed Model y Linear Regression Model fueron las más referenciadas con dos menciones cada una; y los modelos matemáticos más referenciados fueron Modified Abrams-Strogatz y Model of Nonlinear Differential Equations of Reaction-Diffusion Type con tres menciones cada una. Concluimos destacando los estudios de Che et al (2018) y Dwivedi et al (2020) como investigaciones representativas en el uso del aprendizaje de máquina para la identificación de lenguas en peligro de extinción.

Palabras Claves: Aprendizaje automático, Lengua en peligro de extinción, Factores de vitalidad, Predicción, Modelos matemáticos.

Abstract

Languages in danger of extinction are a problem that is occurring rapidly due to various factors. According to United Nations (2023), more than 43% of languages in the world were in danger of extinction. Identifying the vitality of a language is a complicated and important task to maintain cultural diversity, and because it will depend on many causes and the accessibility of data, it may not be entirely predictable. This task of predicting the vitality of a language can be more efficient and accurate thanks to machine learning techniques. In this sense, a systematic review of the literature on language extinction prediction models was carried out, basing the search on five research questions using the PRISMA methodology. This scientific study retrieved 962 articles from the Scopus, IEEE and ScienceDirect databases. The number of studies used to answer the research questions was 33.

Finally, analyzing the machine learning models, the Generalized Linear Mixed Model and Linear Regression Model were the most referenced with two mentions each; and the most referenced mathematical models were Modified Abrams Strogatz and Model of Nonlinear Differential Equations of Reaction-Diffusion Type with three mentions each. We conclude by highlighting the studies by Che et al (2018) and Dwivedi et al (2020) as representative investigations in the use of machine learning for the identification of endangered languages.

Keywords: Machine learning, Endangered language, Vitality factors, Prediction, Mathematical models.

INTRODUCCIÓN

El peligro de la extinción de lenguas originarias en el Perú ha ido en progresivo aumento en las últimas décadas. Según La Sociedad Peruana de Derecho Ambiental SPDA (2010), 29 de las 48 lenguas originarias peruanas se encuentran en peligro de extinción, entre las lenguas en peligro, de acuerdo al Atlas de Lenguas en Peligro en el Mundo de la UNESCO, se puede nombrar: Achuar, Campa Caquinte, Candoshi, Cashibo-Cacataibo, Cashinahua, Chayahuita, Culina, Ese eja, Harakmbut, Huitoto, Machiguenga, Quechua ancashino, Quechua huanuqueño, Quechua ayacuchano, Shipibo-Conibo, Siona / Secoya, Ticuna, Yagua y Yine. De ahí surge la tarea de prevenir la perdida de una lengua nativa en nuestro país y proteger las lenguas en peligro de extinción. Gracias a la gran aceptación de los modelos de machine learning durante los últimos años, ha habido iniciativas por desarrollar modelos para la resolución del problema actual sobre la predicción de lenguas en peligro de extinción, aunque estos estudios no son comunes, se han realizado estudios de modelamiento matemático, gran parte de ellos alineados a la competencia entre dos o mas lenguas en una comunidad.

Por este motivo, el objetivo de este trabajo es analizar la literatura más actualizada relacionada con el tema mediante una revisión sistemática.

La contribución investigativa de este artículo es revisar varios modelos de pronóstico de vitalidad de una lengua desarrollados durante la década 2014-2023 e identificamos factores de vitalidad de la lengua empleados en los estudios. Por otro lado, también revisamos tecnología utilizados en el procesamiento, métricas, medidas e indicadores evaluadores de los modelos y/o métodos de predicción.

El presente artículo se estructura de la siguiente manera: En la primera parte se presenta los materiales y métodos. Dentro, el artículo incluye la metodología de búsqueda, criterios de inclusión y exclusión, preguntas de investigación y selección de estudios. Luego, la siguiente sección muestra los resultados y discusión; incluye la respuesta de cada pregunta de investigación y fuentes de datos. Posteriormente, se presenta un análisis de los estudios incluyendo información adicional para cada pregunta de investigación. Finalmente, el artículo presenta la conclusión y las referencias.

MATERIALES Y MÉTODOS

En esta investigación aplicamos la metodología PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) para ayudarnos a documentar la revisión sistemática de la literatura.

Metodología de búsqueda

La búsqueda por palabras clave se realizó en tres bases de datos científicas de alto impacto: Scopus, IEEE y ScienceDirect, las cuales fueron seleccionadas basadas en la accesibilidad y admisión de consultas avanzadas. La primera búsqueda en la base de datos IEEE incluyó las siguientes palabras claves: ("All Metadata": Language Endangerment) OR ("All Metadata": Endangered Languages) OR ("All Metadata": Endangered Language) AND ("All Metadata": Machine Learning) se filtro las publicaciones de los últimos 9 años, 2014 - 2023, encontramos 60 artículos. La segunda búsqueda en la base de datos ScienceDirect incluimos las palabras claves: "endangered languages" or "Endangered Language" or "Language Endangerment" con filtro de los últimos 9 años, 2014 - 2023, encontramos 38 artículos. La tercera búsqueda en la base de datos Scopus incluimos las palabras claves ("Language Endangerment" OR "Endangered Languages") AND ("Machine Learning" OR "Predicting" OR "Model") con filtro de los últimos 9 años y encontramos 120 artículos. Realizamos una cuarta búsqueda en Scopus, incluimos los documentos relacionados al documento "Predicting Language Endangerment: A Machine Learning Approach" con filtro de los últimos 9 años, 2014 - 2023, y con áreas relacionadas a "Computer Science", "Mathematics", "Engineering", "Decision Science" y encontramos 744 documentos relacionados; en total en Scopus encontramos 864. Las cuatro búsquedas suman un total de 962 artículos. Posteriormente, los datos importantes se extrajeron de características específicas.

Criterios de inclusión y exclusión

A partir de los artículos obtenidos, aplicamos los criterios de inclusión y exclusión para asegurar que fueran importantes para nuestro tema de investigación.
Los artículos elegidos fueron incluidos si cumplían con los siguientes criterios de elegibilidad:

- Modelado/Predicción/Peligro de extinción del lenguaje o términos relacionados en el título, palabras claves o resumen.
- El articulo incluía información sobre el modelado de la vitalidad de un lenguaje.
- Artículos que mostraran información para responder las preguntas de investigación sobre el tema.

Los criterios de exclusión que aplicamos son los siguientes:

- Artículos duplicados en diferentes bases de datos.
- Artículos más enfocados en ciencias sociales.
- Artículos que no mostraran información útil para responder las preguntas de investigación sobre el tema.

Preguntas de investigación

Las preguntas de investigación presentadas en la revisión de la literatura ayudarán a recuperar características importantes de los artículos que podrían aportar información crucial para aplicar

al problema sobre la predicción de la vitalidad de los lenguajes en peligro de extinción en el Perú.

- R1: ¿Qué modelos de machine learning, algoritmos y métodos se utilizan para la predicción de lenguas en peligro de extinción?
- R2: ¿Qué modelos matemáticos se utilizan para la predicción de lenguas en peligro de extinción?
- R3: ¿Qué métricas, medidas e indicadores se utilizan para evaluar los modelos y/o métodos de predicción?
- R4: ¿Qué tecnologías se utilizan para el procesamiento de datos, obtención de datasets y construcción del modelo predictivo?
- R5: ¿Qué factores influyen en la vitalidad de una lengua?

Selección de Estudios

Siguiendo los pasos que indica la metodología PRISMA en la búsqueda con palabras claves, se obtuvieron los documentos más importantes siendo un total de 962 artículos. En primer lugar, se eliminaron las investigaciones duplicadas, por lo que se excluyó 15 artículos. Posteriormente, los 947 artículos pasaron a una segunda evaluación. Considerando el título, resumen y contenido, 880 artículos fueron retirados por no cumplir con los criterios de inclusión y exclusión. La mayoría de los artículos incluidos debían incluir temas relacionados con los modelos de extinción de una lengua o competencia lingüística. Los artículos excluidos fueron en su mayoría de temas relacionados, pero no relevantes, más enfocados en ciencias sociales. Después de aplicar los criterios de elegibilidad, 67 artículos se encontraban en la tercera etapa (50 Scopus, 10 IEEE, 7 ScienceDirect). Debido al análisis anterior, algunos de los artículos no cumplieron con los criterios de inclusión. Por último, 33 fue el número de artículos que se utilizaron para responder a las preguntas de investigación. El proceso de selección según la metodología PRISMA se representa en la **Figura 1**.

RESULTADOS Y DISCUSIÓN

La revisión sistemática incluye 33 artículos sobre modelos de predicción de lenguas en peligro de extinción, modelos matemáticos de competencia lingüística y factores de vitalidad de una lengua. Los artículos procedían de las tres fuentes siguientes: 28 artículos de Scopus (84.84% de los artículos), 2 artículos de IEEE (6.06% de los artículos), 3 artículos de ScienceDirect (9.09% de los artículos).

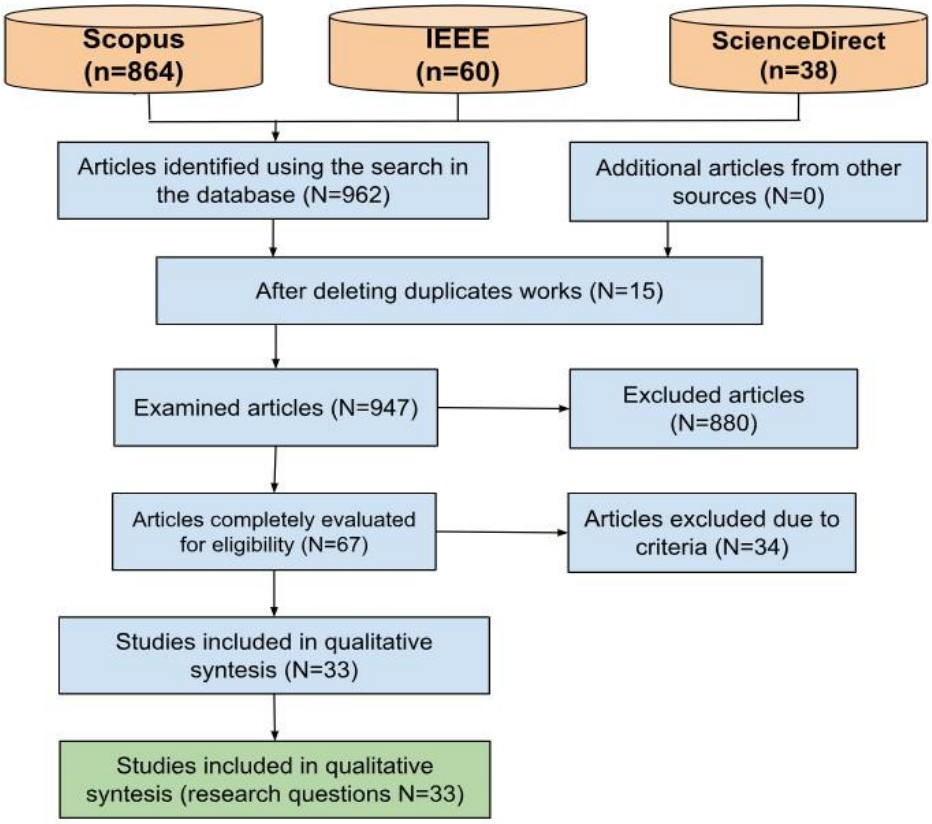

Figura 1. Diagrama utilizando metodología PRISMA para los estudios incluidos en la revisión sistemática.

Respuesta a RQ1

La primera pregunta indaga sobre qué modelos de machine learning, algoritmos y métodos existentes se utilizan para predecir lenguas en peligro de extinción. Los modelos de machine learning, algoritmos y métodos más comunes se muestran en la **Tabla 1** que incluyen en el caso de métodos: Numerical simulations method y Fixed point method con 4 menciones cada uno; en el caso de modelos: Linear Regression Models y Generalized Linear Mixed Model (GLMM) con 2 menciones cada uno. En la **Tabla 1** se puede ver que los artículos utilizaron modelos, métodos y algoritmos en el desarrollo de su investigación.

Tabla 1: Artículos relacionados

Modelo, Algoritmos y Métodos	**Artículos Relacionados**
Linear Regression Model	Dwivedi et al (2020); Amano et al (2014);
Logistic Regression Model	Ikoba and Jolayemi (2020);
Support Vector Machines (SVM)	Che et al (2018);
K-Nearest Neighbors (KNN)	Che et al (2018);
Linear Discriminate Analysis (LDA)	Che et al (2018);
Quadratic Discriminant Analysis (QDA)	Che et al (2018);
Decision Tree	Che et al (2018);
Bagging Method	Che et al (2018);

Random forest	Che et al (2018);
AdaBoosting	Che et al (2018);
Generalized Linear Mixed Model (GLMM)	Bromham et al (2020); Kik et al (2021))
Generalized Linear Model	Amano et al (2014);
Segmented Regression Models	Amano et al (2014);
Quadratic Polynomial Regression Model	Amano et al (2014);
Ordinary Least-Squares (OLS) Model	Amano et al (2014);
The Least-Squares Regression	Isern and Fort (2014);
Spatial Autoregressive (SAR) Model	Amano et al (2014);
Spatial Error Model (SEM)	Amano et al (2014);
Null Regression Models	Amano et al (2014);
Decision Curve Analysis (DCA)	Bromham et al (2020);
Approximate Bayesian Computation (ABC) Method	Zhou et al (2020);
Spectral Method	Eslahchi and Esmaili (2020);
Fourier Spectral Algorithm	Owolabi and Gómez-Aguilar (2018);
Numerical Simulations	Sofuoglu (2017); Nie et al (2015); Owolabi and Gómez-Aguilar (2018); Tchendjeu et al (2020);
Fixed Point Method	Eslahchi and Esmaili (2020); Colucci et al (2014); Colucci et al (2016); Isern and Fort (2014);

Respuesta a RQ2

La segunda pregunta investiga sobre los modelos matemáticos predictivos sobre la extinción de una lengua y modelos de competición lingüística. La **Tabla 2** muestra los modelos matemáticos. Los modelos matemáticos más comunes son el Modified Abrams–Strogatz Model y Model of Nonlinear Differential Equations of Reaction-Diffusion Type con 3 menciones cada una en la **Tabla 2**. En la **Tabla 2** se puede ver que 19 artículos utilizaron modelos de competencia para 2 lenguas y 5 utilizaron modelos de competencia entre más de 2 lenguas.

Tabla 2: Artículos relacionados

Modelos	**Artículos Relacionados**
Modelo de Competencia entre dos Lenguas	
Templin Model	Gazzola and Templin (2022); Templin (2019);
Mathematical Model of Nonlinea Systems of Differential Equations	Díaz and Switkes (2021);
Mathematical Model with Compartmental Epidemiological Modeling Approach	Tchendjeu et al (2020);

Analytical Model of Population Dynamics Based on System of Differential Equations	Seoane et al (2019);
Differential Equation Model	Luck and Mehta (2020);
Model of Nonlinear Differential Equations of Reaction-Diffusion Type	Walters (2014); Cherniha and Davydovych (2020); Cherniha and Davydovych (2021);
Coupled Nonlinear Parabolic Equations Model	Eslahchi and Esmaili (2020);
Reaction-Diffusion Differential Equation Model	Kandler and Unger (2017); Isern and Fort (2014);
Fractional Order Differential Equations Model	Sofuoglu (2017);
Modified Abrams–Strogatz Model	Colucci et al (2014); Colucci et al (2016); Gong et al (2014);
Bilingual Competence Control Model Based on Differential Equations	Nie et al (2015);
Differential Equation Model Combining The Minett-Wang Model and A Simplified Version of The Baggs-Freedman Model	Heinsalu et al (2014);
Social Computational Model Based on The Abrams-Strogatz and Baker Model	Yun et al (2015);
Modelo de competencia entre más de dos lenguas	
Extended Abrams-Strogatz Model	Zhou et al (2020); Paekivi and Rekker (2020);
Model of Nonlinear Differential Equations of The Reaction-Diffusion Type Through The Extended Abrams-Strogatz Model	Owolabi and Gómez-Aguilar (2018);
Naming Game Model With Mobile Agents	Lipowska and Lipowski (2017);
Dynamic Social Network Model	Qi et al (2015);

Respuesta a RQ3

La tercera pregunta investiga sobre que métricas, medidas e indicadores se utilizan para evaluar los modelos y/o métodos de predicción. La **Tabla 3** muestra las métricas, medidas e indicadores. El indicador más común es Equilibrium Points con 8 menciones en la **Tabla 3**.

Tabla 3: Artículos relacionados

Métricas, Medidas e Indicadores	**Artículos Relacionados**
R-Squared (R^2)	Dwivedi et al (2020); Bromham et al (2020);
Recall	Che et al (2018);
Precision	Che et al (2018);

F1 score	Che et al (2018);
Accuracy	Che et al (2018);
ROC curve	Che et al (2018);
Confusion Matrix	Che et al (2018);
Adjusted R^2	Bromham et al (2020);
Spearman R	Kik et al (2021);
Delta Corrected Akaike Information Criterion	Kik et al (2021);
Degrees of Freedom	Kik et al (2021);
P-Value	Kik et al (2021); Ikoba and Jolayemi (2020); Bromham et al (2020);
Z-Value	Amano et al (2014);
coefficients	Amano et al (2014);
Standard Errors	Amano et al (2014);
Pesos de Akaike (wi)	Amano et al (2014);
Chi-Cuadrado	Ikoba and Jolayemi (2020);
Pseudo-R^2*deCoxySnell*	Ikoba and Jolayemi (2020);
Nagelkerke R2	Ikoba and Jolayemi (2020);
Coeficiente de Contingencia de Pearson	Ikoba and Jolayemi (2020); Acharyya and Mahanta (2019);
Standard Deviation	Yun et al (2015); Seoane et al (2019);
Likelihood Ratio	Bromham et al (2020);
Relative Rate	Bromham et al (2020);
Covariance	Zhou et al (2020);
Equilibrium Points	Díaz and Switkes (2021); Tchendjeu et al (2020); Paekivi and Rekker (2020); Owolabi and Gómez- Aguilar (2018); Sofuoglu (2017); Nie et al (2015); Walters (2014); Heinsalu et al (2014);
Coeficientes de Correlación de Rango Parcial (PRCC)	Tchendjeu et al (2020);
Uniform Distribution	Luck and Mehta (2020);
Exponential Distribution	Luck and Mehta (2020);
Average Lifetime	Lipowska and Lipowski (2017);
Uniform Density	Lipowska and Lipowski (2017);

Respuesta a RQ4

La cuarta pregunta indaga sobre las tecnologías que se utilizan para el procesamiento de datos, base de datos para la obtención de datasets y tecnologías utilizadas en la elaboración del modelo predictivo. Las tecnologías presentadas fueron los siguientes: database Glottolog, database ELCat, database UNESCO, database E20, database WLMS, Ethnologue database, database ILD, database idescat, web application GlottoVis, web application GlottoScope, MATLAB, Maple 12, Python y paquetes o librerías de Python como Scikit Learn, astroABC, R y paquetes de R como segmented, spdep, MuMIn, Leafle, Bootstrap v3, Bootstrap sampling. La tecnología más común fue Matlab y Python que se utilizaron en 3 referencias cada una y en el caso de las bases de datos la más común fue Ethnologue con 3 menciones, que se muestran en la **Tabla 4**.

Tabla 4: Artículos relacionados

Tecnologías	**Artículos Relacionados**

Python	Gazzola and Templin (2022); Che et al (2018); Zhou et al (2020);
Leaflet	Dwivedi et al (2020); Hammarstrom et al (2018);
MATLAB	Díaz and Switkes (2021); Eslahchi and Esmaili (2020); Owolabi and Gomez-Aguilar (2018);
Maple 12 Software	Owolabi and Gómez-Aguilar (2018); Cherniha and Davydovych (2020);
Web Application GlottoScope	Hammarstrom et al (2018);
Web Application GlottoVis	Hammarstrom et al (2018);
Glottolog Database	Hammarstrom et al (2018);
ELCat Database	Hammarstrom et al (2018);
Ethnologue Database	Hammarstrom et al (2018); Kik et al (2021); Yun et al (2015);
UNESCO Database	Hammarstrom et al (2018);
E20 Database	Hammarstrom et al (2018);
The World Language Mapping System (WLMS) Database	Amano et al (2014);
Index of Linguistic Diversity (ILD) Database	Amano et al (2014);
Idescat Database	Seoane et al (2019);
Bootstrap V3	Hammarstrom et al (2018);
Bootstrap Sampling	Che et al (2018);
R	Amano et al (2014);

Respuesta a RQ5

La Quinta pregunta se refiere a qué factores influyen en la vitalidad de una lengua y son utilizados por los modelos para pronosticar el peligro de extinción de una lengua. La **Tabla 5** muestra los factores influyentes en la vitalidad de una lengua. Entre los factores más comunes mencionados en la literatura citaban a Number of speakers of a language con 9 menciones, Language Status con 7 menciones, Socioeconomic, Sociocultural y Mortality rate con 5 menciones cada una, que se muestran en la **Tabla 5**.

Tabla 5: Artículos relacionados

Factores de Vitalidad	**Artículos Relacionados**
Sociopolitics	Dressler (2018); Zhou et al (2020); Dwivedi et al (2020); Sofuoglu (2017);
Government and Institutional Language Attitudes and Policies	Acharyya and Mahanta (2019);
Generation	Bromham et al (2020); Ikoba and Jolayemi (2020);
	Dressler (2018); Dwivedi et al

Sociocultural	(2020); Zhou et al (2020); Gong et al (2014); Kandler and Unger (2017);
Sociopsychological	Dressler (2018);
Attitudes of Community Members Towards their Own Language	Acharyya and Mahanta (2019);
Sociolinguistics	Dressler (2018); Heinsalu et al (2014);
Intergenerational Linguistic Transmission	Acharyya and Mahanta (2019); Gazzola and Templin (2022); Templin (2019); Amano et al (2014);
Socioeconomic	Dressler (2018); Dwivedi et al (2020); Kik et al (2021); Zhou et al (2020); Sofuoglu (2017);
Urbanization of The Birth House	Kik et al (2021);
Childhood Place	Ikoba and Jolayemi (2020);
Language Use at Home	Kik et al (2021); Ikoba and Jolayemi (2020);
Community	Bromham et al (2020);
Education Level	Bromham et al (2020); Ikoba and Jolayemi (2020);
Language Learning in Formal Education	Gazzola and Templin (2022); Templin (2019); Tchendjeu et al (2020);
Heritage Language Exposure	Bromham et al (2020);
Materials for Education and Linguistic Literacy	Acharyya and Mahanta (2019);
Quantity and Quality of Documentation	Acharyya and Mahanta (2019);
Number of Speakers of a Language	Dwivedi et al (2020); Kik et al (2021); Colucci et al (2014); Colucci et al (2016); Walters (2014); Templin (2019); Yun et al (2015); Acharyya and Mahanta (2019); Amano et al (2014);
Population Dynamics	Gazzola and Templin (2022); Templin (2019);
Proportion of Speakers of the Language	Qi et al (2015); Seoane et al (2019);
Population of the Region	Walters (2014); Amano et al (2014); Dwivedi et al (2020); Qi et al (2015);
Trends in Existing Linguistic Domains	Acharyya and Mahanta (2019);
Response to New Domains and Media	Acharyya and Mahanta (2019);
Gross Domestic Product (GDP)	Che et al (2018); Amano et al (2014);
Latitude and Longitude	Che et al (2018);
Countries in Which a Language is Spoken	Che et al (2018);
Global Peace Index	Che et al (2018);
Human Freedom Index	Che et al (2018);
World Risk Index	Che et al (2018);
Greenbergs Diversity Index	Che et al (2018);
Family Size	Che et al (2018);
Language Learning for Adults	Gazzola and Templin (2022); Templin (2019);
Parents' Language Skills	Kik et al (2021);
Migration	Gazzola and Templin (2022); Lipowska and Lipowski (2017);
Immigration	Díaz and Switkes (2021);
Linguistic Richness	Amano et al (2014);
Language Status	Eslahchi and Esmaili (2020); Kandler and Unger (2017); Isern and Fort (2014); Kik et al (2021); Qi et al (2015);

	Díaz and Switkes (2021); Heinsalu et al (2014);
Social Interaction (Social Radio)	Qi et al (2015); Díaz and Switkes (2021); Sofuoglu (2017);
Recruitment Rate	Tchendjeu et al (2020);
Mortality Rate	Tchendjeu et al (2020); Heinsalu et al (2014); Yun et al (2015); Owolabi and Gómez-Aguilar (2018); Yun et al (2015);
Birth Rates	Owolabi and Gómez-Aguilar (2018); Yun et al (2015);
Rate of Contact Between Individuals in the Community about Language	Tchendjeu et al (2020);
Language Forgetting Factor	Tchendjeu et al (2020); Yun et al (2015);
Demographic Dynamics	Luck and Mehta (2020); Kandler and Unger (2017); Amano et al (2014);
Population Mobility	Eslahchi and Esmaili (2020); Qi et al (2015);
Population Growth	Owolabi and Gómez-Aguilar (2018); Eslahchi and Esmaili (2020); Amano et al (2014);
Initial Frequency Distributions	Kandler and Unger (2017);
Linguistic Composition	Templin (2019);
Dynamics and Linguistic Environment	Templin (2019);
Family Formation	Templin (2019);
State-Dependent Impulsive Control Strategies	Nie et al (2015);
Annual Precipitation	Amano et al (2014);
Vegetation Productivity	Amano et al (2014);
Temperature Seasonality	Amano et al (2014);
Precipitation Seasonality	Amano et al (2014);
Elevation Range	Amano et al (2014);
Habitat Diversity	Amano et al (2014);

Fuentes de datos

La **Tabla 6** indica el tipo de fuente (Journal papers, Conference proceedings, Books) de las diferentes bases de datos. La mayoría de las publicaciones provinieron de Scopus (28 estudios). Además, la **Tabla 6** muestra que la mayoría de los estudios provienen de Journal papers (30 en total) con el 90.9% de las investigaciones referenciadas.

Tabla 6: Diferente tipo de fuente por base de datos.

Tipo	Scopus	IEEE	ScienceDirect	Total	Porcentaje
Conference Proceedings	0	2	0	2	**6.06%**
Journal Papers	27	0	3	30	**90.9%**
Books	1	0	0	1	**3.03%**
Total	**28**	**2**	**3**	**33**	**100%**

ANALISIS DE LOS ESTUDIOS

El Además de responder las preguntas, se realiza un análisis adicional para cada pregunta de investigación. En las siguientes subsecciones se presentan cifras que indican el conteo de referencias por pregunta de investigación y algunos comentarios importantes.

Modelos, Algoritmos y/o Métodos (Q1)

Un importante aspecto para mencionar es que hay estudios recurrentes que usan Linear Regression Models que es una técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras; Modelo mixto lineal generalizado (GLMM) que amplía el modelo lineal de modo que el objetivo está linealmente relacionado con los factores y covariables mediante una función de enlace especificada; Numerical simulations method que es un método de investigación por computadora que ejecuta cálculos basados en un modelo matemático específico para simular procesos físicos reales; y Fixed point method que también se conoce como método de iteración simple de punto fijo, en el cual se utiliza una fórmula para predecir la raíz de una función, la misma que puede desarrollarse por una iteración simple.

En la **Figura 2** un gráfico de barras representa el número de menciones por Modelos, Algoritmos y/o Métodos según la **Tabla 1**.

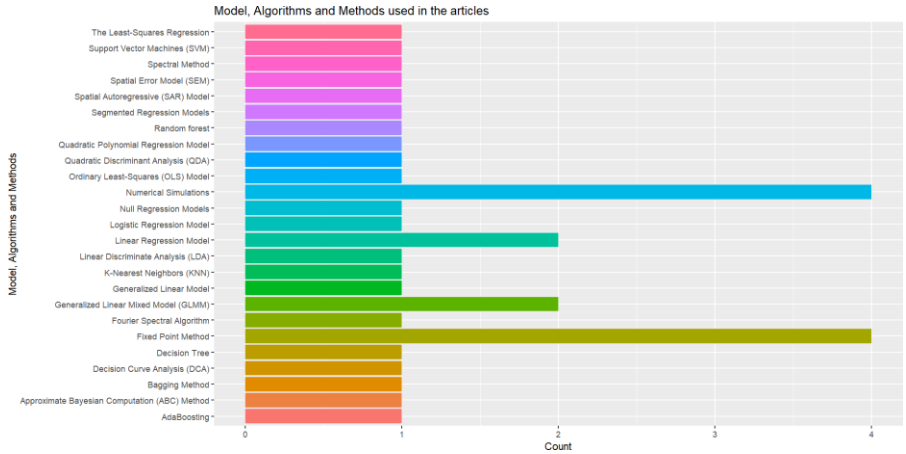

Figura 2. Conteo de Modelos, Algoritmos y/o Métodos en la revisión sistemática.

Modelos Matemáticos (Q2)

Los modelos matemáticos mas citados fue el Model of Nonlinear Differential Equations of Reaction-Diffusion Type y el Modified Abrams–Strogatz Model en este caso ambos modelos estudian la competencia entre dos lenguas. En la **Figura 3** un gráfico de barras representa el número de citaciones de los modelos matemáticos de competencia lingüística según la **Tabla 2**.

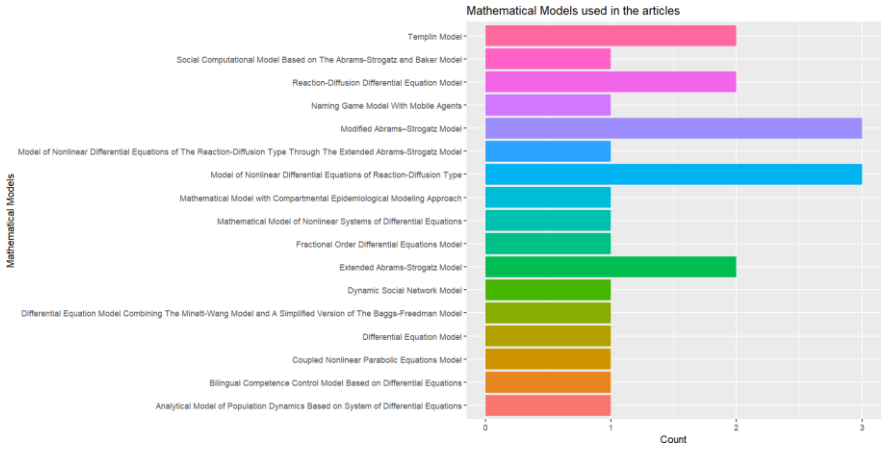

Figura 3. Conteo de Modelos Matemáticos en la revisión sistemática.

Métricas, Medidas e Indicadores (Q3)

Las métricas fueron usadas por los autores para medir el rendimiento de un modelo de aprendizaje automático y fue solo una elección de los autores decidir que métricas usar. También se incluyen medidas estadísticas en los diferentes estudios científicos. Sin embargo, el indicador más esencial fue Equilibrium Points incluido en la mayoría de los artículos con 8 menciones. En la **Figura 4**, un gráfico de barras representa el número de menciones por métrica, medidas e indicadores según la **Tabla 3**.

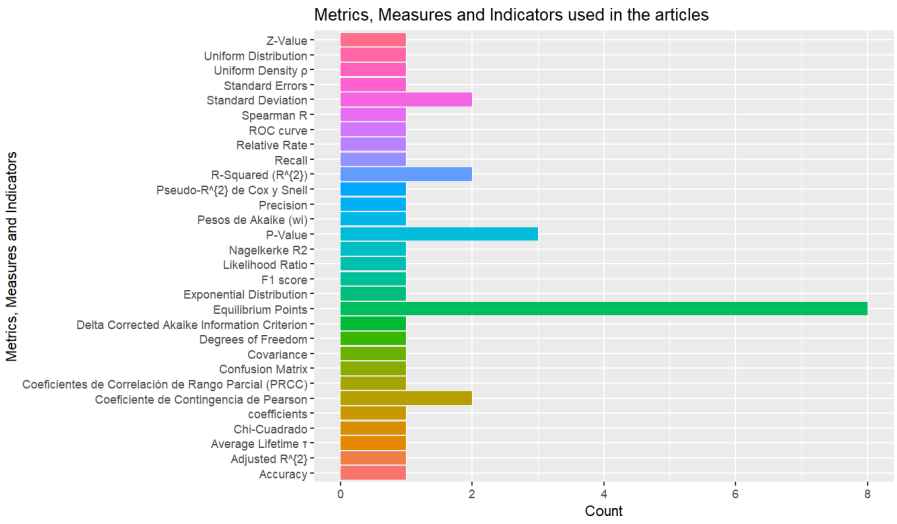

Figura 4. Conteo de Métricas, Medidas e Indicadores en la revisión sistemática.

Tecnologías (Q4)

Podemos mencionar que usan Matlab de una forma reiterativa, este software es una plataforma de programación y cálculo numérico para el desarrollo de algoritmos, análisis de datos, visualización y cálculo numérico. Además, esto se debe a que la mayoría de artículos analizados trabajaron con modelos matemáticos en el análisis de la vitalidad de una lengua. Así también se usa Python en algunos artículos que trabajaron con los modelos de machine learning. La Base de datos más usada fue Ethnologue Database donde se puede encontrar, leer e investigar las más de 7.000 lenguas vivas conocidas en el mundo, y obtener acceso a estadísticas y otra

información sobre las lenguas vivas del mundo. En la **Figura 5** un gráfico de barras representa el número de menciones por tecnologías según la **Tabla 4**.

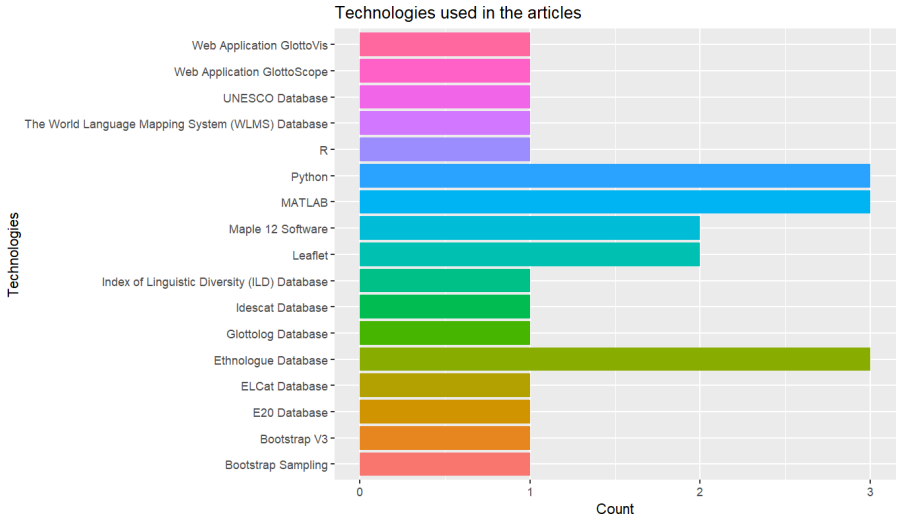

Figura 5. Conteo de Tecnologías en la revisión sistemática.

Factores de Vitalidad (Q5)

Los factores de vitalidad mas citados fueron Number of Speakers of a Language, Language Status, Mortality Rate, Socioeconomic, Sociocultural, Sociopolitics, Population total of the Region y Intergenerational Linguistic Transmission. Los factores más importantes mencionados en la **Tabla 5** fueron Number of Speakers of a Language (9 artículos) y Language Status (7 artículos). Es necesario mencionar que los niveles Sociopolitics, Sociocultural, Sociopsychological, Sociolinguistics y Socioeconomic, se pueden considerar factores generales que incluyen otros factores mencionados en la **Tabla 5**, pero se está interpretando cada término por separado, teniendo en cuenta que no todos los factores se pueden asociar a niéveles. En la **Figura 6** un gráfico de barras representa el número de menciones por factor de vitalidad según la **Tabla 5**.

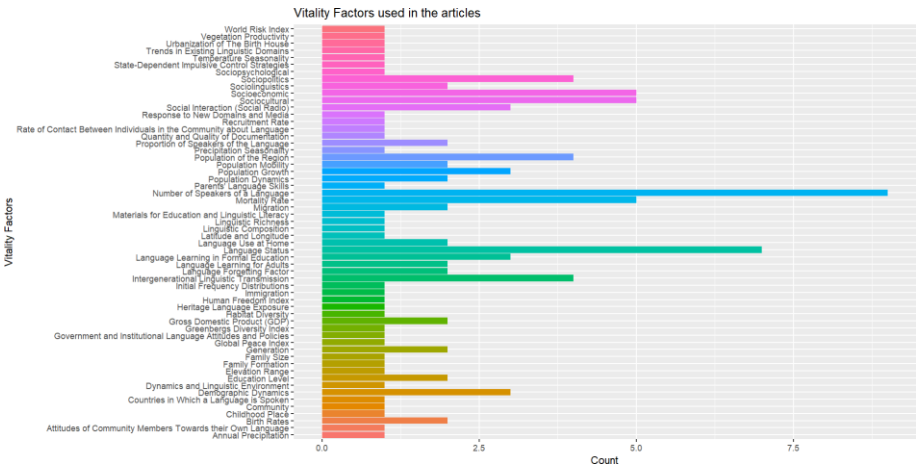

Figura 6. Conteo de Factores de Vitalidad en la revisión sistemática.

CONCLUSIONES

El Este articulo presento una revisión sistemática de la literatura de 962 artículos relacionados a los modelos de predicción de lenguas en peligro de extinción que se basaron en métodos de Inteligencia Artificial y modelamiento matemáticos de competencia, en la que se revisaron los resúmenes de estudios para finalmente obtener 33 artículos relevantes para este artículo. La revisión se realizó mediante el estudio y análisis de artículos científicos publicados en bases de datos Scopus, IEEE y ScienceDirect. Los artículos más importantes se revisaron según las cinco preguntas de investigación presentadas en el Marco General y los criterios de elegibilidad. Los investigadores han mostrado interés en estudiar modelos de predicción de extinción de una lengua y la mayoría de los estudios incluyeron factores de vitalidad, métricas, tecnologías y base de datos para implementar los modelos en sus estudios. La literatura analizada presenta el potencial del aprendizaje automático y el modelamiento matemático para la detección de una lengua en peligro de extinción. Este estudio, permite trazar los factores que influyen en la vitalidad de una lengua, es necesario mencionarlo ya que todos los estudios se centran en uno o más factores para evaluar la vitalidad. De los 33 artículos que fueron analizados cabe resaltar el estudio desarrollado por Dwivedi et al (2020) que presenta un modelo de aprendizaje automático que pronostica una tendencia decreciente y un posible cronograma de extinción de una lengua, el modelo utilizado fue Linear Regression Model que si bien no fue el mayor citado, este estudio desempeña un papel crucial para marcar un antes y después respecto a la predicción de vitalidad de una lengua con modelos de regresión, y complementando con los resultados de la pregunta 5 (RQ5) sobre factores que influyen en la vitalidad de una lengua, podemos incluir otros factores para trabajos futuros. Asimismo, otra investigación importante fue el de Che et al (2018) quien introduce el aprendizaje automático para predecir la extinción de una lengua, este estudio utiliza modelos de clasificación. Con lo mencionado concluimos que el uso de machine learning está surgiendo en los últimos años para identificar lenguas en peligro de extinción y ayudar a prevenir la perdida de las mismas.

REFERENCIAS BIBLIOGRÁFICAS

Acharyya P, Mahanta S (2019) Language vitality assessment of Deori: An endangered language. Language Documentation and Conservation 13:514– 544

Amano T, Sandel B, Eager H, et al (2014) Global distribution and drivers of language extinction risk. Proceedings of the Royal Society B: Biological Sciences 281(1793):17–19. https://doi.org/10.1098/rspb.2014.1574

Bromham L, Hua X, Algy C, et al (2020) Language endangerment: A multidimensional analysis of risk factors. Journal of Language Evolution 5(1):75–91. https://doi.org/10.1093/jole/lzaa002

Che D, Shafer T, Tian P (2018) Classification of endangered languages using decision tree based algorithms. ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery pp 1814–1821. https://doi.org/10.1109/FSKD.2017.8393042

Cherniha R, Davydovych V (2020) Exact solutions of a mathematical model describing competition and coexistence of different language speakers. Entropy 22(2):1–11. https://doi.org/10.3390/e22020154

Cherniha R, Davydovych V (2021) On a Nonlinear Mathematical Model for the Description of the Competition and Coexistence of Different-Language Speakers. Journal of Mathematical Sciences (United States) 256(5):628–639. https://doi.org/10.1007/s10958-021-05449-5

Colucci R, Mira J, Nieto JJ, et al (2014) Coexistence in Exotic Scenarios of a Modified Abrams–Strogatz Model. COMPLEXITY https://doi.org/10.1002/cplx.21623

Colucci R, Mira J, Nieto JJ, et al (2016) Non Trivial Coexistence Conditions for a Model of Language Competition Obtained by Bifurcation Theory. Acta Applicandae Mathematicae 146(1):187–203. https://doi.org/10.1007/ s10440-016-0064-3, URL http://dx.doi.org/10.1007/s10440-016-0064-3

Díaz M, Switkes J (2021) Speaking out: A mathematical model of language preservation. Heliyon 7(5):e06,975. https://doi.org/10.1016/j.heliyon.2021. e06975, URL https://doi.org/10.1016/j.heliyon.2021.e06975

Dressler WU (2018) Independent, Dependent and Interdependent Variables in Language Decay and Language Death. European Review 26(1):120–129. https://doi.org/10.1017/S1062798717000370

Dwivedi P, Shraddha C, Mathews S, et al (2020) Predicting Language Endangerment: A Machine Learning Approach. 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020 https://doi.org/10.1109/ICCCNT49239.2020.9225576

Eslahchi MR, Esmaili S (2020) The convergence and stability analysis of a numerical method for solving a mathematical model of language competition. Applied Numerical Mathematics 151:119–140. https://doi.org/10.1016/j.apnum.2019.12.015, URL https://doi.org/10.1016/j.apnum.2019.12. 015

Gazzola M, Templin T (2022) Language Competition and Language Shift in Friuli-Venezia Giulia: Projection and Trajectory for the Number of Friulian Speakers to 2050. Sustainability (Switzerland) 14(6). https://doi.org/ 10.3390/su14063319

Gong T, Shuai L, Zhang M (2014) Modelling language evolution: Examples and predictions. Physics of Life Reviews 11(2):280–302. https://doi.org/10.1016/j.plrev.2013.11.009, URL http://dx.doi.org/10.1016/j.plrev.2013.11.009

Hammarström H, Castermans T, Forkel R, et al (2018) Simultaneous visualization of language endangerment and language description. Language Documentation and Conservation 12(July):359–392

Heinsalu E, Patriarca M, Léonard JL (2014) The role of bilinguals in language competition. Advances in Complex Systems 17(1):1–16. https://doi.org/10.1142/S0219525914500039

Ikoba NA, Jolayemi ET (2020) Investigation of Factors Contributing to Indigenous Language Decline in Nigeria. Philippine Statistician 69(2):55–70

Isern N, Fort J (2014) Language extinction and linguistic fronts. Journal of the Royal Society Interface 11(94):2–10. https://doi.org/10.1098/rsif.2014.0028

Kandler A, Unger R (2017) Modeling language shift. Diffusive Spreading in Nature, Technology and Society pp 351–373. https://doi.org/10.1007/ 978-3-319-67798-9¯18

Kik A, Adamec M, Aikhenvald AY, et al (2021) Language and ethnobiological skills decline precipitously in Papua New Guinea, the world's most linguisti- cally diverse nation. Proceedings of the National Academy of Sciences of the United States of America 118(22). https://doi.org/10.1073/pnas.2100096118

Lipowska D, Lipowski A (2017) Language competition in a population of migrating agents. Physical review E 95(5-1):052,308. https://doi.org/10.1103/PhysRevE.95.052308, https://arxiv.org/abs/arXiv:1702.07888

Luck JM, Mehta A (2020) On the coexistence of competing languages. European Physical Journal B 93(4):1–18. https://doi.org/10.1140/epjb/ e2020-10038-1, https://arxiv.org/abs/arXiv:2003.04748

Nie LF, Teng ZD, Nieto JJ, et al (2015) State impulsive control strategies for a two-languages competitive model with bilingualism and interlinguistic similarity. Physica A: Statistical Mechanics and its Applications 430(11461067):136–147. https://doi.org/10.1016/j.physa.2015.02.064, URL http://dx.doi.org/10.1016/j.physa.2015.02.064

Owolabi KM, Gómez-Aguilar JF (2018) Numerical simulations of multilingual competition dynamics with nonlocal derivative. Chaos, Solitons and Fractals 117:175–182. https://doi.org/10.1016/j.chaos.2018.10.020, URL https://doi.org/10.1016/j.chaos.2018.10.020

Paekivi S, Rekker A (2020) Modeling the competition between three language groups. AIP Conference Proceedings 2302(October 2017). https://doi.org/10.1063/5.0033498

Qi YP, Ya L, Bi GH, et al (2015) Complex agent network model for the competition among three languages. Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015 pp 5910–5917. https://doi.org/10.1109/CCDC.2015.7161868

Seoane LF, Loredo X, Monteagudo H, et al (2019) Is the coexistence of Catalan and Spanish possible in Catalonia? Palgrave Communications 5(1):1–9. https://doi.org/10.1057/s41599-019-0347-1, URL http://dx.doi.org/10.1057/s41599-019-0347-1, https://arxiv.org/abs/arXiv:1801.08117

Sofuoglu Y (2017) Ordinary and fractional mathematical models on language competition and bilingualism. Applied Sciences 19(1):122–131

SPDA W (2010) Unesco: 29 lenguas originarias peruanas permanecen en peligro de extinción. https://www.actualidadambiental.pe/ unesco-29-lenguas-originarias-peruanas-permanecen-en-peligro-de-extincion/, accessed: 2023

Tchendjeu AE, Bowong S, Tchitnga R, et al (2020) Dynamics of the competition between two languages. SeMA Journal 77(4):351–373. https://doi.org/10.1007/s40324-020-00219-w, URL https://doi.org/10.1007/s40324-020-00219-w

Templin T (2019) A language competition model for new minorities. Rationality and Society 31(1):40–69. https://doi.org/10.1177/1043463118787487

United Nations (2023). Día Internacional de la Lengua Materna | Naciones Unidas. https://www.un.org/es/observances/mother-language-day, accessed: 2023

Walters CE (2014) A reaction–diffusion model for competing languages. Meccanica 49(9):2189–2206. https://doi.org/10.1007/s11012-014-9973-2

Yun J, Li XT, Liu S, et al (2015) Social computational model of language endangerment and recovery. Open Cybernetics and Systemics Journal 9(1):1524–1529. https://doi.org/10.2174/1874110X01509011524

Zhou Z, Szymanski BK, Gao J (2020) Modeling competitive evolution of multiple languages. PLoS ONE 15(5):1–16. https://doi.org/10.1371/journal.pone.0232888, URL http://dx.doi.org/10.1371/journal.pone. 0232888, https://arxiv.org/abs/arXiv:1907.06848