Machine Learning meets Linguistics: A case of classification of endangered languages

Project 4

Group 10 - Amina MAHFOUD/Gabriel OSMAN/ Ammar SAMAD KHAN

Table of content

01

Introduction & Project Overview

- Introduction and objective of the project
- Reason for our choice of project topic
- Research Questions
- Approach

02

Data Model Implementation

- Data cleanup and Analysis
- Initialize, train and evaluate a model
- Utilize retrieved data and demonstrate accuracy

03

Data Model Optimization

- Model optimization and evaluation
- Display revised model performance

04

Conclusion

- Overall summary
- Languages and Libraries deployed
- Ethical considerations and references

01. Introduction & Project Overview

Project Overview

This project aims to investigate the most widely spoken languages and those at risk of extinction using machine learning techniques. By analyzing the prevalence of languages, we can shed light on the factors contributing to language vitality and identify opportunities for language preservation.

Why this topic?

We consider the topic of language has not been given the needed attention. It was an opportunity to study the history of languages spoken across the world, identify how the ones close to us that are rarely heard of or spoken, are dying away.
Could it be because of modernization or immigration, overall an interesting topic not discussed enough.

Objective

- To analyze and visualize the current status of the world's most widely spoken languages and those at risk of extinction
- To employ machine learning models to predict the category of languages based on their attributes and obtain an accuracy score that can correctly classify languages that face extinction.

Research Question

- Can machine learning accurately predict the category of a language using degree of endangered?

01. Introduction & Project Overview (continued)

Approach

Supervised Machine Learning

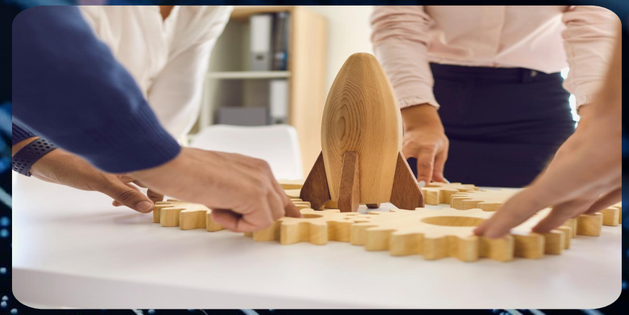
We decided on Supervised Learning since our project objective was to accurately classify spoken languages around the world into levels or degrees of endangered, for this we need:

- Structured and Labelled Data
- Scalable Data
- Performance metrics for comparisons

Classification Models

This predicts the category the data belongs to:

- Decision boundaries that separate different classes, i.e degrees of endangerment.
- Labelled data utilization
- Performance measurement and optimization using metrics like Accuracy, precision, recall and F1 score



Models Used:

- **Logistic Regression,**
- **XGBoost and**
- **Random Forest**

02. Data Model Implementation

Data cleanup and Analysis

- Extract dataset in csv from data sources, raw dataset had over 2,000 rows, 15 columns
- Transformed/Processed the raw dataset by dropping, renaming and processing NaN values.
- Load the cleaned data set into saved csv data files.
- Exploratory Analysis by plotting the spread of endangered languages over a world map (**demo 1**).

Initialize, train and evaluate a model

- From exploration of the data, we identify required Features and Targets as X and Y
- Split the dataset into training and testing sets.
- Train using XGBoost, Logistics regression and Random Forest Classifier
-

Utilize retrieved data and demonstrate accuracy

- Run and read the accuracy score
- Initial accuracy score were as follows:
  - Logistic regression - 40%
  - XGBoost - 49%
  - Random Forest 47%

Next we considered optimization steps

03. Data Model Optimization

Changes made to the data:

- Using beautifulsoup scraped additional data about Languages such as most spoken languages datasets, website usages, languages learned on duolingo and population growth.
- Performed additional cleaning of data by cleaning, merging, dropping and filling NaN values.

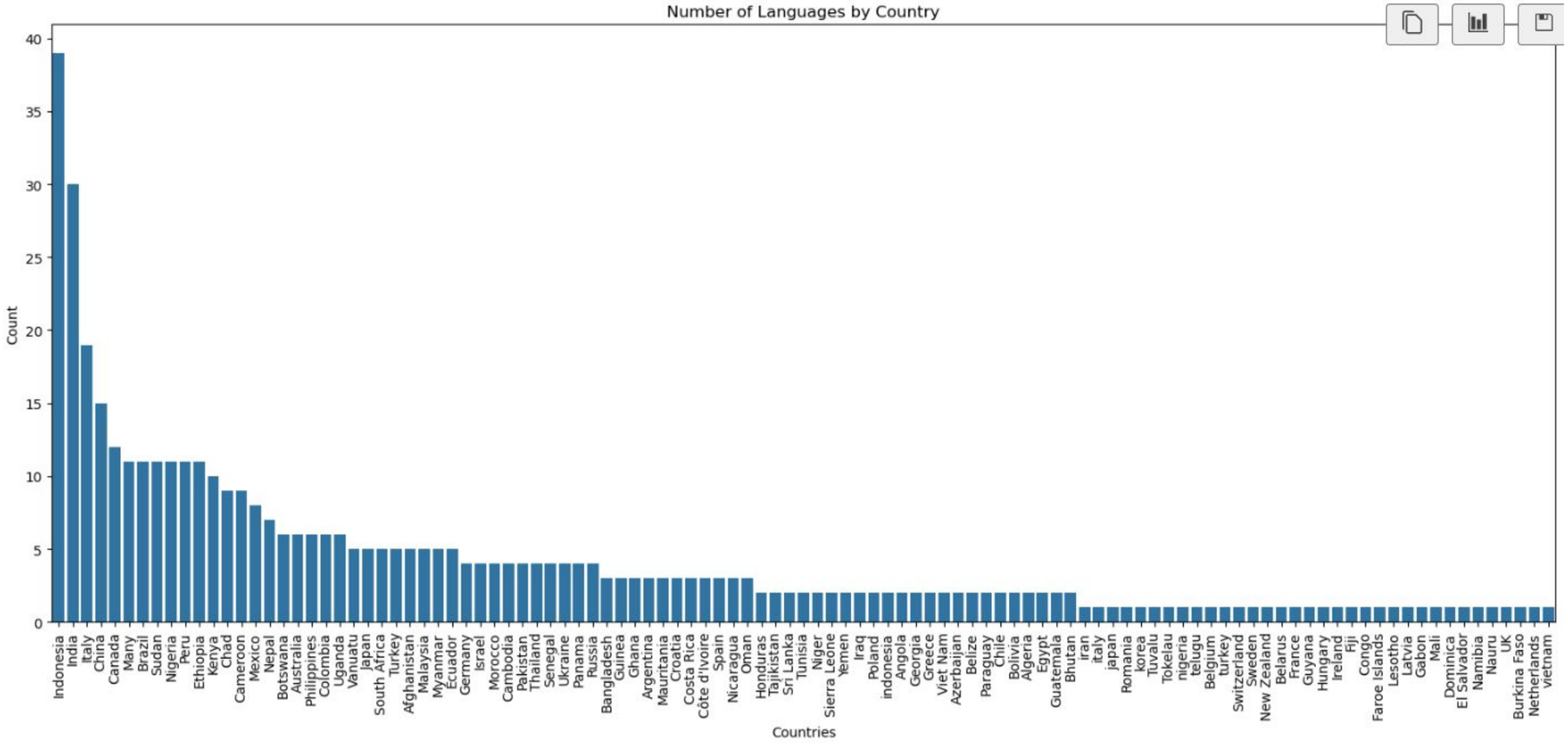
Further optimization using code:

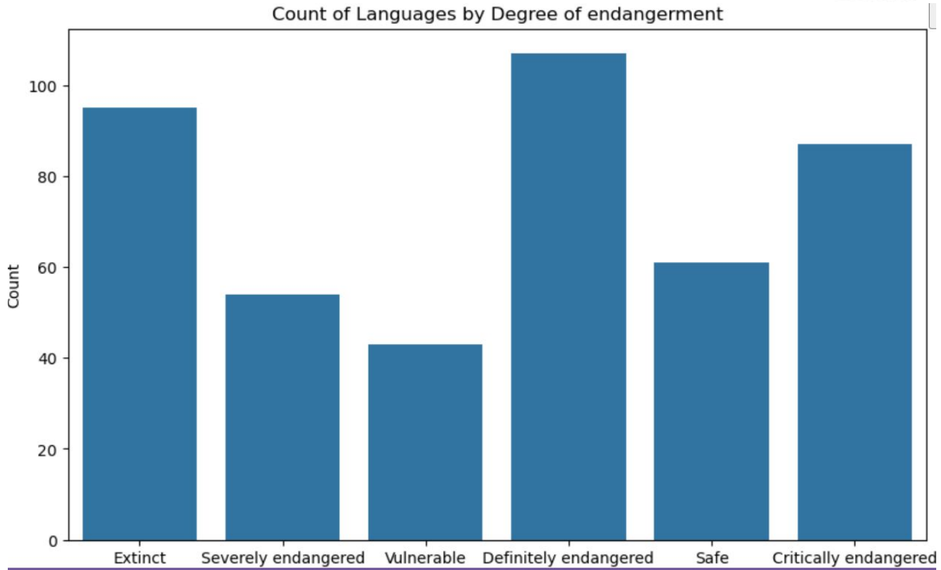
- Loaded the revised dataset with iterative changes into a csv, rerun the model using RandomForest
- Evaluating the model by displaying accuracy using a Confusion matrix, review importance of Features
- further optimizing model using GridSearchCV code, print and review results in a classification report.

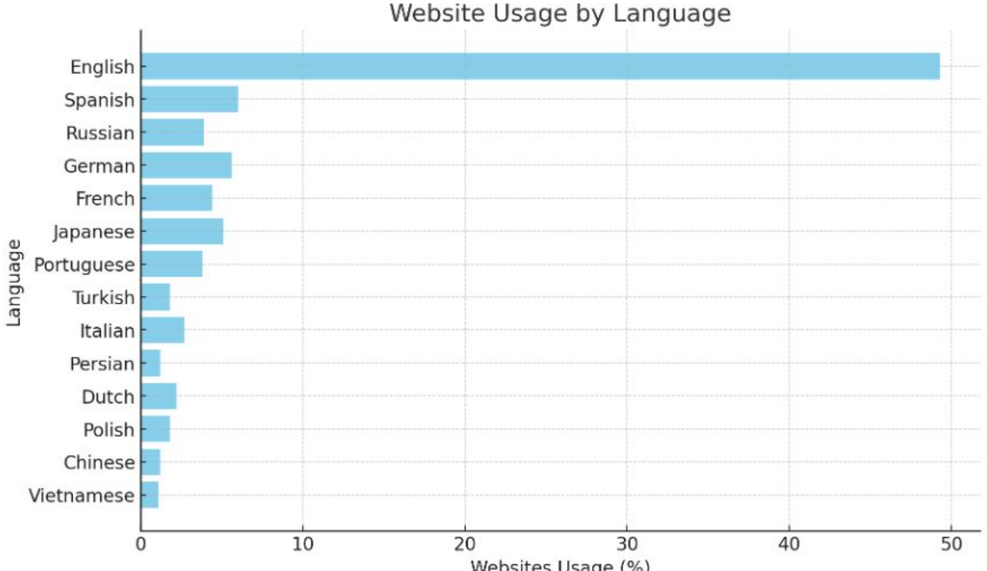
Overall model performance post optimization

The accuracy score increase from between 40% and 50%, to 75% representing the following:

- A classification model on largely non-binary or numeric data requires more data to properly train a model.
-


Number of Languages by Country
Count
Countries
Indonesia
India
Italy
China
Canada
Many
Brazil
Sudan
Nigeria
Peru
Ethiopia
Kenya
Chad
Cameroon
Mexico
Nepal
Botswana
Australia
Philippines
Colombia
Uganda
Vanuatu
Japan
South Africa
Turkey
Afghanistan
Malaysia
Myanmar
Ecuador
Germany
Israel
Morocco
Cambodia
Pakistan
Thailand
Senegal
Ukraine
Panama
Russia
Bangladesh
Guinea
Ghana
Argentina
Mauritania
Croatia
Costa Rica
Côte d'Ivoire
Spain
Nicaragua
Oman
Honduras
Tajikistan
Sri Lanka
Tunisia
Niger
Sierra Leone
Yemen
Iraq
Poland
indonesia
Angola
Georgia
Greece
Viet Nam
Azerbaijan
Belize
Paraguay
Chile
Bolivia
Algeria
Egypt
Guatemala
Bhutan
iran
italy
japan
Romania
korea
Tuvalu
Tokelau
nigeria
telugu
Belgium
turkey
Switzerland
Sweden
New Zealand
Belarus
France
Guyana
Hungary
Ireland
Fiji
Congo
Faroe Islands
Lesotho
Latvia
Gabon
Mali
Dominica
El Salvador
Namibia
Nauru
UK
Burkina Faso
Netherlands
vietnam

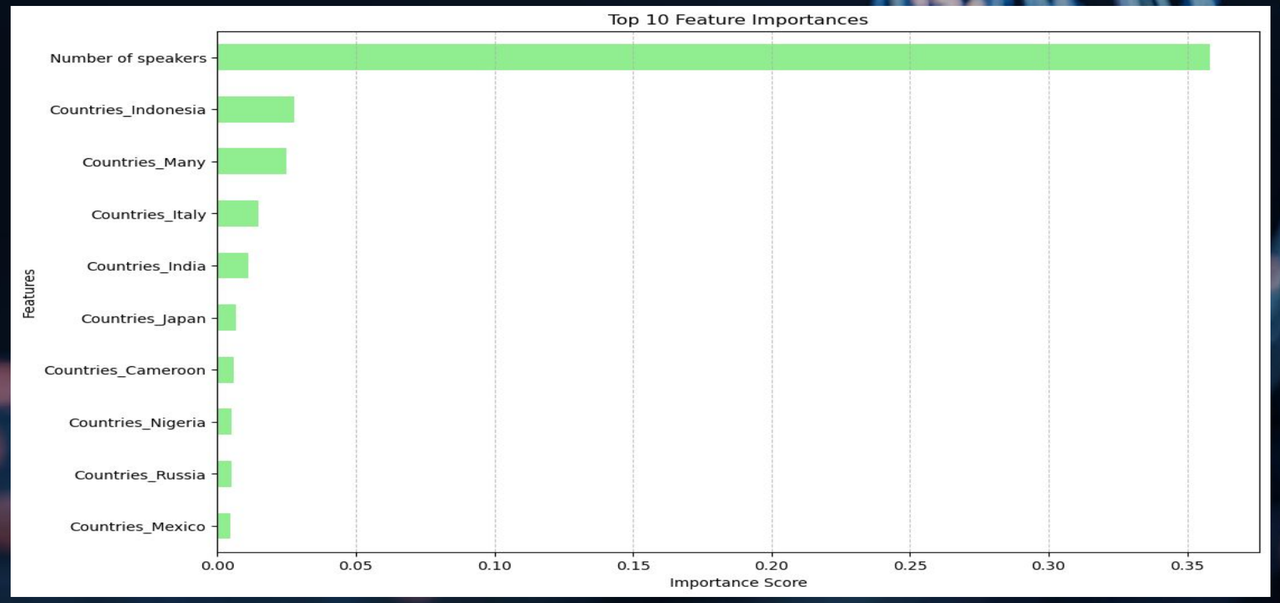

Count of Languages by Degree of endangerment
Count
Extinct
Severely endangered
Vulnerable
Definitely endangered
Safe
Critically endangered


Website Usage by Language
Language
English
Spanish
Russian
German
French
Japanese
Portuguese
Turkish
Italian
Persian
Dutch
Polish
Chinese
Vietnamese
Websites Usage (%)

03. Data Model Optimization (continued)

Accuracy Score : 0.75
Classification Report

	precision	recall	f1-score	support
Critically endangered	0.68	1.00	0.81	23
Definitely endangered	0.65	0.73	0.69	30
Extinct	1.00	1.00	1.00	26
Safe	1.00	1.00	1.00	13
Severely endangered	0.00	0.00	0.00	7
Vulnerable	0.00	0.00	0.00	13
accuracy			0.75	112
macro avg	0.55	0.62	0.58	112
weighted avg	0.66	0.75	0.70	112



Display model Performance

Overall Model Performance:

- Accuracy: The model correctly classified 75% of the cases.
- Macro Avg: Averages precision, recall, and F1-score equally across all classes (useful when class imbalance exists).
- Weighted Avg: Averages the metrics but considers the number of instances per class.

Features Importance:

The "Number of speakers" is the dominant feature, contributing the most to the model's predictions. Country-related features have much lower importance in determining endangerment levels. This suggests the model relies heavily on number of speaker.

04. CONCLUSION

04. CONCLUSION (continued)

Summary

The code effectively demonstrates a machine learning pipeline that can help predict the category of languages at risk of extinction based primarily on the number of speakers.

Language & Libraries:

python Programming SQLite3 Database

pandas

JS JavaScript

HTML CSS

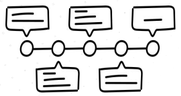
SQL

dmlc XGBoost

Flask web development, one drop at a time

04. CONCLUSION (continued)

- Front end visualization
- Overall summary
- Languages and Libraries deployed
- Ethical considerations and references



References:

- Extinct Languages (https://www.kaggle.com/datasets/the-guardian/extinct-languages/data)
- the guardian article ‘ Endangered Languages: the full list - (https://www.theguardian.com/news/datablog/2011/apr/15/language-extinct-endangered#data)
- https://www.endangeredlanguages.com/

Ethical consideration

Minimize Bias and Misrepresentation:

- Ensure transformations do not mislead users or alter original meanings
- Justify changes in data representation (e.g., aggregating data).
- Continue to ensure integrity with data normalization

Ensure transparency:

- Maintain detailed documentation of the transformations applied.
- Ensure users can trace back to original values if necessary.

Correct citation and references of source of datasets

THANK YOU!

Any Questions?