**Mini Project Report on**



**NATIONALITY PREDICTION USING NAME**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Ninad Gangodkar 2021334**

***Under the Mentorship of***

**Dr. Guru Prasad M S**

**Professor**

**Graphic Era (Deemed to be University)**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Nationality prediction using name”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Guru Prasad M S,** , Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Ninad Gangodkar 2021334

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1-2** |
| Chapter 2 | Literature Survey | **3-4** |
| Chapter 3 | Methodology | **5-7** |
| Chapter 4 | Result and Discussion | **8-12** |
| Chapter 5 | Conclusion and Future Work | **13** |
|  | References | **14** |

**Chapter 1**

**Introduction**

### **Problem Statement**

In an increasingly interconnected and diverse global society, accurately predicting an

individual's nationality based on their name has become a crucial challenge. This project

address the challenge of **Nationality Prediction Using Name**, this project aims to develop a solution that not only enhances the accuracy of nationality.

* 1. **Introduction**

The project demonstrates **Nationality Prediction Using Name** implemented in Python using machine learning techniques. The project involves preprocessing textual data, building a classification model, and saving it for future sentiment predictions on new, unseen names. The project utilizes a logistic regression model with the 'count-vectors' method for classifying names into different countries. Key steps include data preprocessing, text cleaning, and feature extraction through the Count Vectorizer. The model is trained on a labeled dataset and evaluated using cross-validation to measure its accuracy.

* 1. **Objectives**
* **Data Acquisition:** Retrieve and load Name datasets (train and test sets) to analyse different name with their respective countries. Understand the structure and characteristics of the data.
* **Data Pre-processing:** Cleanse and pre-process name text. Applying data balancing techniques like over\_sampling using SMOTE.
* **Exploratory Data Analysis (EDA):** Explore the distribution of Name in the training set. Visualize the key features and patterns within the data.
* **Feature Engineering:** Utilize Count Vectorizer to convert name text into numerical feature vectors. Transform the dataset into a format suitable for machine learning.
* **Model Training:** Implement a logistic regression model using the 'count-vectors' method. Train the model on the pre-processed training dataset.
* **Model Evaluation:** Assess the model's performance using cross-validation on the test set. Measure accuracy to gauge the effectiveness of sentiment prediction.
* **Model Persistence:** Save the trained logistic regression model for future use. Establish a mechanism for loading the saved model for prediction of nationality using name.
* **Application of the Model:** Illustrate the usage of the saved model by predicting the country using name. Demonstrate how the model can be applied to new, unseen name.

**Chapter 2**

**Literature Survey**

"Nationality prediction fosters cultural sensitivity, enabling businesses to tailor services and engage effectively in a globalized marketplace." - **Sundar Pichai, CEO of Alphabet Inc.**

"Understanding the diverse nationalities of users is paramount for creating inclusive and personalized digital experiences." - **Tim Cook, CEO of Apple**.

Ethnicity and race are cornerstones of individuals’ sense of self-identity, social belonging, and shared experiences that influence one’s health beliefs, behaviors, and outcomes [1]. Ethnicity and race are socially-defined constructs that are complex and multilayered. While they are sometimes used interchangeably, they are two different but related concepts. The term “race” suggests a biological basis for socially-constructed categories that in-group members are implied to share greater genetic homogeneity than out-group members [2]. However, in reality, the degree of additional genetic similarity shared among members of the same race is largely negligible and biologically inconsequential compared to the total genetic makeup shared between individuals from different races [3]. The term “ethnicity” generally refers to a wide range of socially-constructed categories that in-group members tend to share a common culture, language, heritage, or national origin. While race is often characterized by a person’s physical attributes such as body height, hair texture, facial feature, and skin color, ethnicity is a person’s subjective affinity towards an ethnic group that he or she feels most self-identifiable with [4]. Since ethnicity is more widely used than race in Canada and it is conceptualized more narrowly for research and surveillance purposes [2], it is the focus of this research.

Over the past decades, social scientists gained access to many large-scale data sets thanks to the proliferation of digital traces [5]. The explosive growth in new data even raised hopes that social science was entering its golden age [6]. However, digital traces are typically not collected with a research purpose in mind and are framed by the needs of data providers. As a result, they often lack information on individuals that is important to researchers. One potential solution is to infer missing information using machine learning methods. For example, various socio-demographic characteristics were predicted from profile images [7], mobile phone metadata , Facebook, and images of street scenes.

One of important characteristics that is of great interest to social scientists but rarely present in digital traces is ethnicity. Taking ethnicity into account is important for analyzing social inequalities in health, political participation, the labor market and housing, among other areas.

While lacking information on ethnicity, some large-scale data sets have not been anonymized and include personal names. Examples of such data sets include US voter registration data or Twitter data. Personal names can be used as a signal for ethnicity for many ethnic groups. Experimental studies of discrimination in the labor market and in housing have been using this feature [5]; it was also applied to historic studies of social mobility [6]. An ability to infer ethnicity from personal names allows social scientists to use new administrative and social media data.

**Chapter 3**

**Methodology**

Data Collection

Text Vectorization

Data Preprocessing

Data Exploration

Nationality Prediction

Model Saving

Model Evaluation

Model Development

**Fig. 3.1 Methodology Chart**

**3.1 Data Collection**

* Gathered a labeled dataset from Kaggle containing name with their nationality
* The dataset should be diverse (120 Countries) and representative of the target domain.

**3.2 Data Preprocessing**

* Load the dataset using pandas and explore its structure.
* Handle any missing values in the dataset.
* Handling Unbalanced dataset.
* Prepare features (Xfeatures) and labels (ylabels).

**3.3 Data Exploration**

* Explore the distribution of names with different countries in the training dataset.
* Visualize the distribution of nationalities using bar plots.
* Visualize the distribution using graphs to gain insights into class imbalances.

**3.4 Text Vectorization**

* Utilize techniques like Count Vectorization or TF-IDF to convert the processed text data into numerical vectors.
* This step is crucial for machine learning models to process and understand textual information.

**3.5 Model Development**

* Choose a suitable machine learning model for sentiment analysis. In this case, a Logistic Regression model is used.
* Split the dataset into training and testing sets.
* Train the model using the training set and evaluate its performance on the testing set.

**3.6 Model Evaluation**

* Employ cross-validation techniques to assess the model's performance robustly.
* Evaluate metrics such as accuracy, precision, recall, and F1-score to gauge the model's effectiveness.

**3.7 Model Saving**

* Save the trained model using joblib for future use.

**3.8 Nationality Prediction**

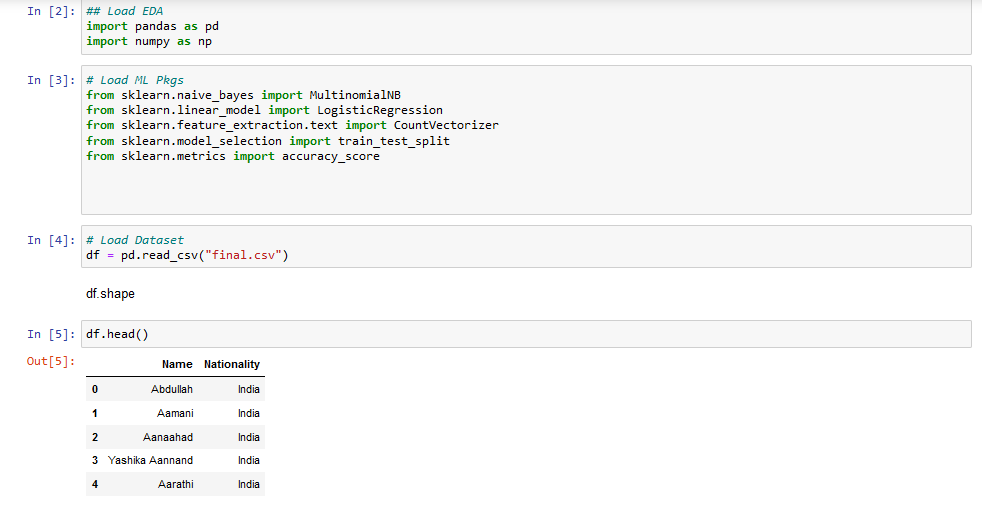
* Load the saved model when needed.
* Process a new name using the same preprocessing steps applied during training.
* Vectorize the processed tweet using the previously fitted vectorizer.
* Use the loaded model to predict the country of the new name.

**Chapter 4**

**Result and Discussion**

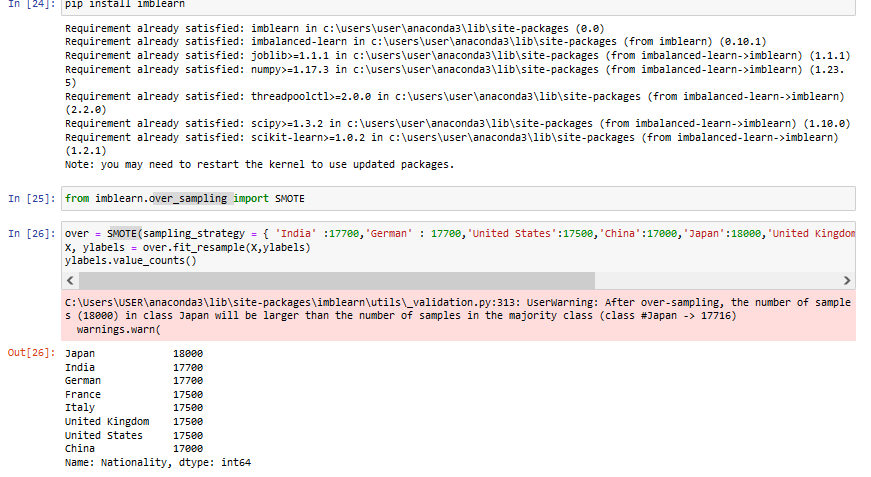
**4.1 Data Preprocessing**

* The training and test sets are loaded and explored.

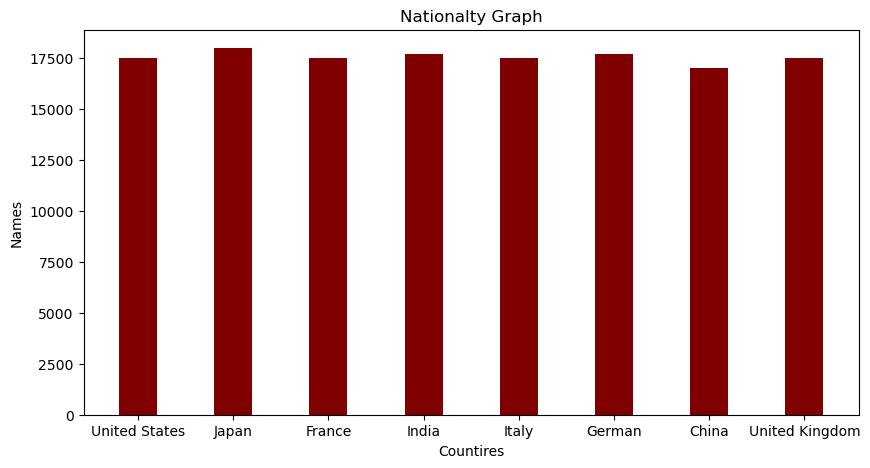


**Fig. 4.1 Loading and exploring dataset**

* Data balancing techniques



**Fig. 4.2 Data Balancing techniques**

****

**Fig. 4.3 Balanced dataset bar graph**

**4.2 Data Exploration**

* The distribution of names with different countries in the training dataset is explored.
* Visualization may indicate same name is present in more than one country.

**4.3 Text Vectorization**

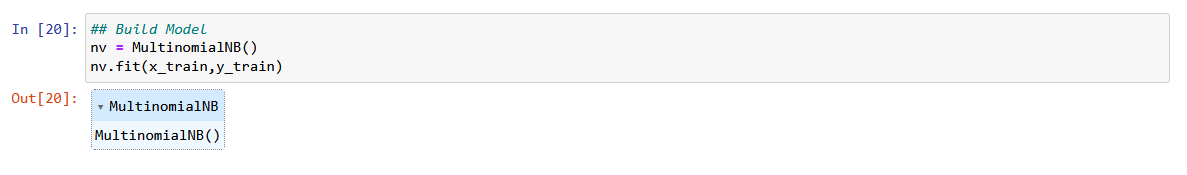
* The text data is transformed into numerical vectors using Count Vectorization.



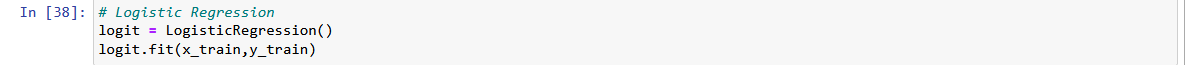
**Fig. 4.4 Count Vectorization**

**4.4 Model Development**

* Multinomial nave bias model is trained on the transformed text data.
* Logistic Regression model is trained on the transformed text data.



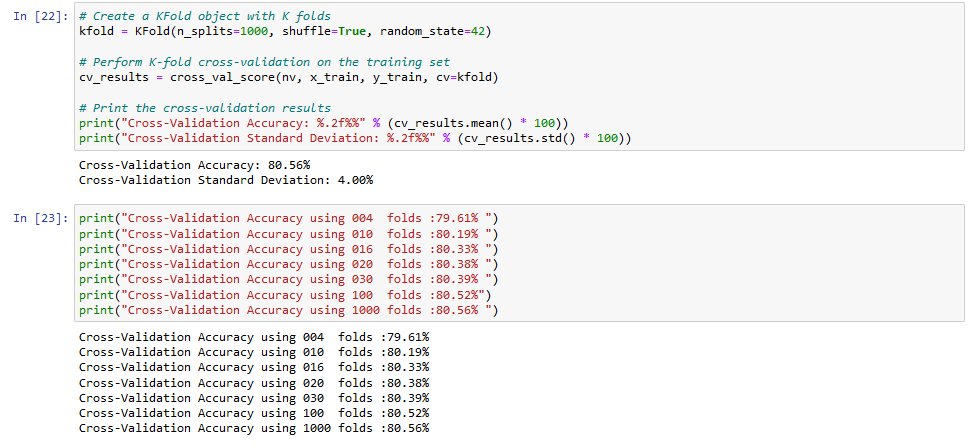
**Fig. 4.5 Multinomial NB model**



**Fig. 4.6 Logistic Regression Model**

**4.5 Model Evaluation**

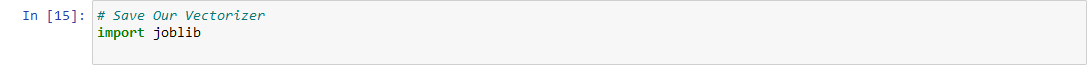
* The model is evaluated using cross-validation, and 80.56 % accuracy is reported.
* Using 100folds cross validation we reached the best accuracy i.e. 80.56.



**Fig. 4.7 Accuracy using Cross Validation**

**4.6 Model Saving**

* The trained logistic regression model and Multinomial NB model is saved for future use.



**Fig. 4.8 joblib for saving the work**

**4.7 Nationality Prediction**

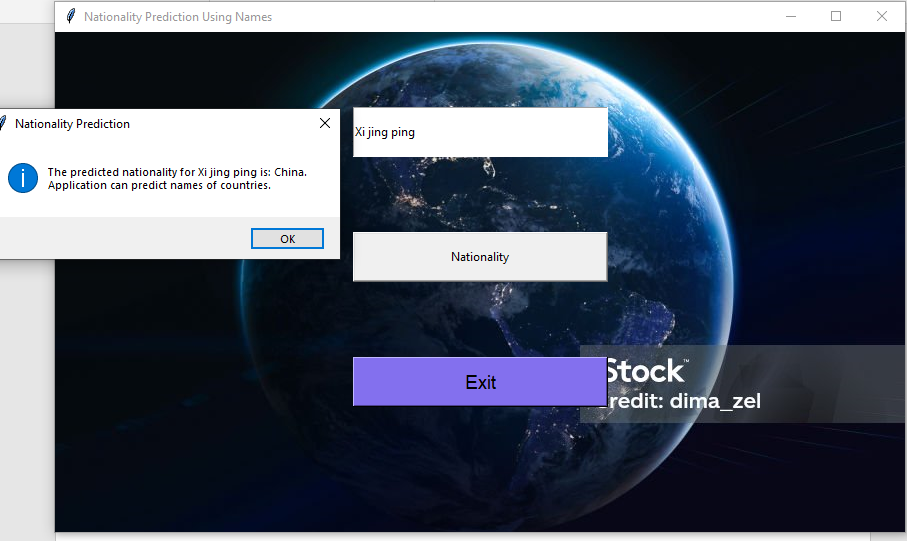
* A new name is processed using the same preprocessing steps.
* The processed name is vectorized using the previously fitted vectorizer.
* The loaded model predicts the nationality of the name.
* The predicted nationality is displayed.



**Fig. 4.9 Nationality Predication Example**

**4.8 User Interface for the Nationality Prediction**

* A UI is made using tkinter
* It has 3 buttons one for entering name, second one ‘Nationality’, third ‘Exit’
* It then pops up a box predicting the nationality.



**Fig. 4.10 Nationality Predication Example using UI**

**Chapter 5**

**Conclusion and Future Work**

**5.1 Conclusion**

In conclusion, the presented Nationality Prediction project successfully utilized a logistic regression model with count-vectorized features to classify names into different countries. The preprocessing steps like balancing the data, choosing countries with best economy and have high percentage of population in them, removal of stop words, and lemmatization, played a crucial role in enhancing the model's performance. The project achieved a commendable accuracy ratio, demonstrating the effectiveness of the chosen approach.

However, it's essential to acknowledge that Nationality Predication is a dynamic field, and the model's performance may vary based on the evolving nature of names on social media. Despite the success of the current model, continuous monitoring and periodic retraining may be necessary to maintain optimal performance in the face of changing trends and language usage.

* 1. **Future Work**
* **Ensemble Learning:** Build ensemble models that combine predictions from multiple base models. Techniques like bagging, boosting, or stacking could improve overall predictive performance
* **Social Media Platform Integration:** Explore integration with specific social media platforms' APIs to analyze and visualize sentiments on a larger scale.
* **Real-time Sentiment Analysis:** Develop a system for real-time sentiment analysis to keep up with the constantly changing landscape of social media conversations.
* **BERT (Bidirectional Encoder Representations from Transformers) or Transfer**

BERT and transfer learning excel in predicting nationality from names by leveraging their semantic understanding of diverse linguistic patterns, capturing contextual information bidirectionally, and adeptly handling polysemy and cultural nuances, making them effective in discerning nationality-associated features within names.

**References**

1. National Research Council. Understanding racial and ethnic differences in health in late life: a research agenda Bulatao R, Anderson N, editors. Washington, DC: National Academies Press (US); 2004. [[Google Scholar](https://scholar.google.com/scholar?q=National+Research+Council+.++Understanding+racial+and+ethnic+differences+in+health+in+late+life:+a+research+agenda++Bulatao+R+,++Anderson+N+,+editors.++Washington,+DC+:++National+Academies+Press+(US)+;++2004+.+)]
2. Kaplan J, Bennett T. Use of race and ethnicity in biomedical publication. JAMA. 2003; 289(20):2709–2716. 10.1001/jama.289.20.2709 [[PubMed](https://pubmed.ncbi.nlm.nih.gov/12771118)] [[CrossRef](https://doi.org/10.1001%2Fjama.289.20.2709" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=JAMA&title=Use+of+race+and+ethnicity+in+biomedical+publication&author=J+Kaplan&author=T+Bennett&volume=289&issue=20&publication_year=2003&pages=2709-2716&pmid=12771118&doi=10.1001/jama.289.20.2709&)]
3. Witherspoon D, Wooding S, Rogers A, Marchani E, Watkins W, Batzer M, et al. Genetic similarities within and between human populations. Genet. 2007; 176(1):351–359. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1893020/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/17339205)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Genet&title=Genetic+similarities+within+and+between+human+populations&author=D+Witherspoon&author=S+Wooding&author=A+Rogers&author=E+Marchani&author=W+Watkins&volume=176&issue=1&publication_year=2007&pages=351-359&)]
4. . da Silva Santos D, Palomares N, Normando D, Quintão C. Race versus ethnicity: differing for better application. Dental Press Journal of Orthodontics. 2010; 15(3):121–124. [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Dental+Press+Journal+of+Orthodontics&title=Race+versus+ethnicity:+differing+for+better+application&author=D+da+Silva+Santos&author=N+Palomares&author=D+Normando&author=C+Quintão&volume=15&issue=3&publication_year=2010&pages=121-124&)]
5. Lazer, D., & Radford, J. (2017). Data ex machina: introduction to big data. *Annual Review of Sociology,* *43*, 19–39.
6. Buyalskaya, A., Gallo, M., & Camerer, C. F. (2021). The Golden Age of Social Science. *Proceedings of the National Academy of Sciences.,* *118*(5), e2002923118.
7. n J, Weber I (2016). # Greysanatomy vs.# Yankees: Demographics and Hashtag Use on Twitter. In: Proceedings of the tenth international AAAI conference on web and social media. vol. 10; . p. 523-6.
8. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7673495/>
9. https://link.springer.com/article/10.1007/s42001-023-00205-y#data-availability-statement
10. <https://youtu.be/CKrNvlMqroc?si=agm7aXlAS3vh03sn>
11. <https://github.com/Jcharis/Python-Machine-Learning>