HANSHIN UNIVERSITY

Hanshin SCAI Research Paper

School of Computing and Artificial Intelligence, Hanshin University

탐색적 데이터 분석 및 머신러닝 응용 프로그램을 통한 인구 추세 예측

Exploratory Data Analysis and Machine Learning Application to Predict Population Trends

Lim Seong soo*

Department of Information and Communication Engineering Hanshin University, Osan-si, Korea

[Abstract]

This research analyzes population dynamics using exploratory data analysis (EDA) techniques and machine learning models to forecast global birth and death trends. By employing datasets from publicly available sources, including Kaggle and government data APIs, we extract meaningful insights on how demographic factors influence population changes. The findings of this research not only provide valuable insights for policymakers but also demonstrate the significance of AI in addressing global challenges like population growth and decline.

Keywords: Exploratory Data Analysis, Machine Learning, Population Trends, Birth and Death Analysis, Data Visualization

Ⅰ. 서 론

세계 인구가 계속 변화함에 따라 출생 및 사망 추세를 이해하는 것은 전세계에 있는 연구자에게 중요한 문제가 되었습니다. 유엔은 세계 인구가 2100년까지 100억 명에 도달할 것으로 추정하며, 이는 자원 배분, 도시 계획 및 환경 관리에 있어 전례 없는 과제를 제시합니다[1]. 이 논문은 과거 인구 데이터를 분석하고 고급 EDA 및 머신러닝 기술을 사용하여 미래 추세를 예측하는 데 중점을 둡니다. data.gov.kr 및 Kaggle을 포함한 여러 소스의 데이터를 통합하고 선형 회귀 및 신경망과 같은 머신 러닝 알고리즘을 활용하였으며, 이 연구는 시간에 따른 인구 변화의 주요 요인에 대한 통찰력을 제공합니다.

Ⅱ. 인구 추세 예측을 위한 python 활용

2-1) python 라이브러리를 설명

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import
LinearRegression
import statsmodels.api as sm
```

1. Pandas: 데이터프레임 형식으로 데이터를 효율적으로 관리하고 조작할 수 있도록 지원합니다. 데이터 정제, 필터링, 그룹화 등 데이터 전처리에 매우 유용합니다.
2. NumPy: 다차원 배열 및 수치 연산을 위한 라이브러리로, 대규모 데이터 처리에 적합합니다. 행렬과 같은 통계적 분석에 자주 사용됩니다.
3. Matplotlib 및 Seaborn: 데이터 시각화를 위한 라이브러리로, 인구 추세의 시각적 표현에 사용됩니다. Matplotlib은 기본적인 그래프와 플롯을 생성하는 데 적합하며, Seaborn은 통계적 플롯과 복잡한 시각화를 쉽게 구현할 수 있습니다.
4. Scikit-learn: 예측 모델을 생성하기 위한 머신러닝 라이브러리입니다. 회귀 분석, 분류, 군집화 등 다양한 알고리즘을 제공하여 인구 추세를 학습하고 예측하는 데 활용됩니다.
5. Statsmodels: 통계 모델링 및 시계열 분석을 지원하는

라이브러리로, 인구 데이터의 시계열 패턴을 분석하고 미래를 예측하는 데 효과적입니다.

2-2) 데이터 전처리 및 데이터베이스 관리

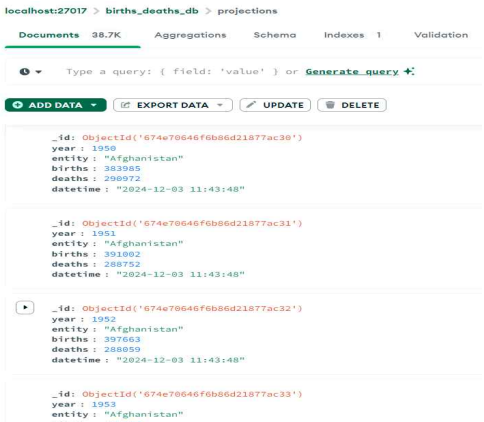

그림 1. MongoDB에 데이터를 저장하는 과정
Fig. 1. How to store data in MongoDB

MongoDB에 데이터를 저장함으로써 다양한 형태의 데이터를 쉽게 저장하고 관리할 수 있습니다.

III. 데이터 구성 및 처리

3-1) 데이터 설명

본 연구에서는 인구 추세 예측을 위해 오픈소스 데이터 플랫폼 Kaggle에서 제공하는 데이터[2]를 활용하였습니다. 해당 데이터는 전 세계 국가별 출생 및 사망 데이터를 포함하고 있으며, Entity, Code, Year, Deaths - Sex: all - Age: all - Variant: estimates, Deaths - Sex: all - Age: all - Variant: medium, Births - Sex: all - Age: all - Variant: estimates, Births - Sex: all - Age: all - Variant: medium의 8개 변수를 포함하고 있습니다.
이 데이터 세트는 총 170여 개의 국가에 대해 1950년부터 2021년까지의 출생 및 사망 데이터를 포함하고 있으며, 다양한 국가별 인구 변화를 분석할 수 있는 기초 데이터를 제공합니다.

3-2) 데이터 구성 방식

	Entity	Code	Year	Deaths - Sex: all - Ag...	Deaths - Sex: all - Ag...	Births - Sex: all - Age:...
1	Afghanistan	AFG	1950	290972		383985
2	Afghanistan	AFG	1951	288752		391002
3	Afghanistan	AFG	1952	288059		397663
4	Afghanistan	AFG	1953	287712		404666
5	Afghanistan	AFG	1954	289189		410428
6	Afghanistan	AFG	1955	286725		417650
7	Afghanistan	AFG	1956	286654		425071
8	Afghanistan	AFG	1957	286113.97		432044
9	Afghanistan	AFG	1958	285911		439867
10	Afghanistan	AFG	1959	285300		447510
11	Afghanistan	AFG	1960	286159		456418
12	Afghanistan	AFG	1961	286818		465920
13	Afghanistan	AFG	1962	288056		476486
14	Afghanistan	AFG	1963	289745		487782
15	Afghanistan	AFG	1964	291215		499081
16	Afghanistan	AFG	1965	292987		510721

그림 2. CSV 데이터 구성방식
Fig. 2. How csv data is organized

각 데이터는 국가 단위로 연도별로 나열되어 있으며, 특정 시점에서 출생 및 사망 데이터를 비교하여 추세를 분석할 수 있는 구조로 설계되어 있습니다.

3-3) 데이터 활용 및 처리

데이터 전처리 과정은 다음과 같습니다.

1. 결측값 처리: 일부 열에서 누락된 값이 발견되었으며, 결측값은 Pandas 라이브러리의 dropna() 또는 fillna() 메서드를 사용하여 처리하였습니다.
2. 변수 정리: 데이터 분석에 필요하지 않은 변수는 제거하고, 주요 변수(Entity, Year, Births, Deaths)만 남겼습니다.
3. 데이터 형 변환: 출생 및 사망 데이터가 문자열로 저장된 경우 이를 정수형으로 변환하여 계산과 시각화에 용이하도록 하였습니다.
4. 추세 분석을 위한 연산: 국가별 연도별로 출생자 수와 사망자 수의 차이를 계산하여 자연 증가율을 계산하였습니다.

IV. 관련 연구

4-1) 탐색적 데이터 분석(EDA)

EDA는 종종 통계적 요약과 시각화를 사용하여 패턴, 이상치 및 추세를 식별하기 위해 데이터를 조사하는 것을 포함합니다[3]. 이 연구에서는 Matplotlib 및 Seaborn과 같은 도구를 사용하여 데이터의 시각적 표현을 만들고 출산율, 사망률 및 경제 지표와 같은 인구 통계 변수 간의

숨겨진 상관관계를 밝혀냅니다.

4-2) 머신러닝 응용 분야

머신러닝은 선형 회귀, 랜덤 포레스트, 신경망과 같은 알고리즘이 시계열 데이터 예측에서 높은 정확도를 보임에 따라 인구 추세를 예측하는 데 점점 더 많이 적용되고 있습니다[4]. 이 연구에서 머신러닝 모델은 모델 해석 가능성과 성능 평가에 초점을 맞춰 과거 패턴을 기반으로 미래의 출생률과 사망률을 예측하도록 훈련됩니다.

Ⅴ. 구현 및 결과

5-1) 선형 회귀 모델을 활용한 데이터 시각화

선형 회귀 모델을 사용해 데이터의 구조와 주요 통계적 특징을 시각화하고, 데이터 간의 관계를 파악하는 작업을 수행했습니다. 본 연구에서는 지역별 출생률과 사망률을 분석하여 중요한 추세를 시각화하고, 이를 통해 해당 지역들의 인구 변화 추세를 식별했습니다.

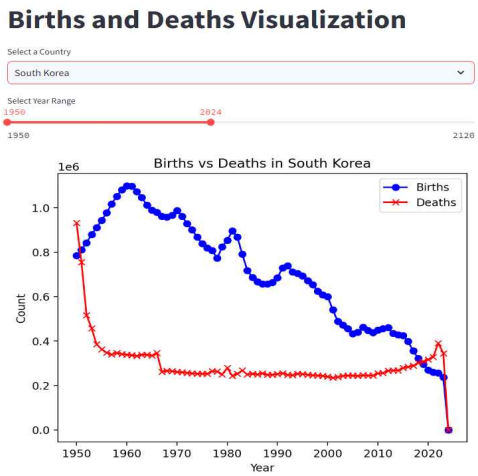

그림 3. 한국의 출산율과 사망률 시각화
Fig. 3. Visualization of fertility and mortality rates in Korea

위 그림은 한국의 출산율과 사망률을 시각화한 그래프로 출산율은 매년 감소하고 있는 반면에 사망률은 유지되고 있는 점을 확인할 수 있다.

그림 4. 일본의 출산율과 사망률 시각화

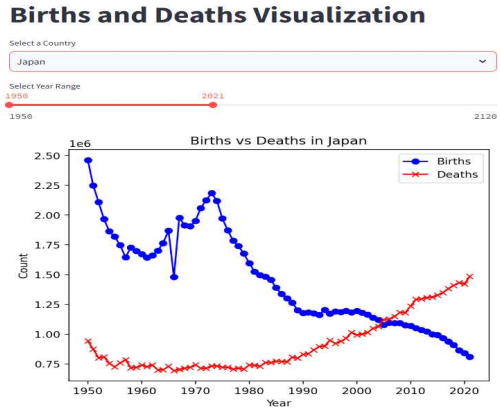

그림 4. 일본의 출산율과 사망률 시각화
Fig. 4. Visualization of fertility and mortality rates in Japan

위 그림은 일본의 출산율과 사망률을 시각화한 그래프로 출산율은 매년 감소하고 사망률이 늘어가는 점을 확인할 수 있다.

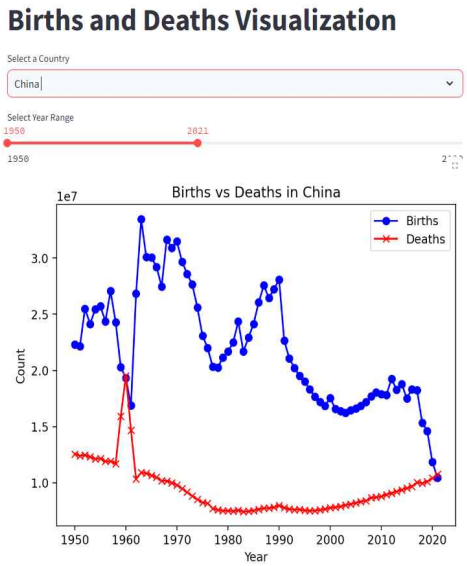

그림 5. 중국의 출산율과 사망률 시각화
Fig. 5. Visualization of fertility and mortality rates in China

위 그림은 중국의 출산율과 사망률을 시각화한 그래프로 출산율은 매년 감소하고 있는 반면에 사망률은 유지되고 있는 점을 확인할 수 있다.

5-2) 선진국의 추세

그림3, 그림4, 그림5를 확인해 알 수 있는 공통적인 특징은 출산율이 매년 감소하고 있다는 점으로 1970년대부터 선진국들은 경제적 발전, 사회적 변화, 그리고 여성의 교육 수준 향상 등 여러 요인으로 출산율이 감소하는 경향을 보였습니다. 이는 특히 선진국에서 더욱 두드러졌습니다. 또 다른 특징으로는 사망률이 안정되었다는 점입니다. 사망률은 선진국에서 안정적으로 유지되거나 점진적으로 감소하는 경향을 보였습니다. 이는 건강 관리, 의료 기술의 발전, 생활 환경의 개선 등으로 인한 것입니다.

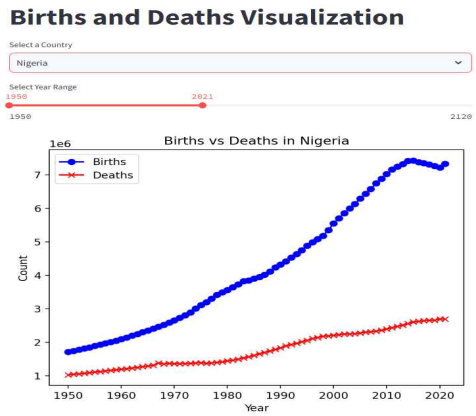

그림 6. 나이지리아의 출산율과 사망률 시각화
Fig. 6. Visualization of fertility and mortality rates in Nigeria

위 그림은 나이지리아의 출산율과 사망률을 시각화한 그래프로 출산율은 현재까지도 매년 증가한 반면 사망률은 유지되고 있는 점을 확인할 수 있다.

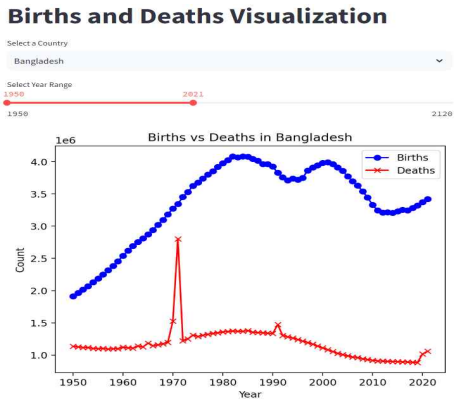

그림 7. 방글라데시의 출산율과 사망률 시각화
Fig. 7. Visualization of fertility and mortality rates in Bangladesh

위 그림은 방글라데시의 출산율과 사망률을 시각화한 그래프로 출산율은 현재까지도 매년 증가했지만, 사망률은 유지되고 있는 점을 확인할 수 있다.

5-3) 개발도상국의 추세

그림6, 그림7을 통해 확인할 수 있는 공통적인 특징은 모두 출산율이 높은 편을 유지하고 있습니다. 두 국가는 급속한 산업화와 경제적 성장을 이루었지만, 사회적·문화적 요인 때문에 출산율 감소 속도는 상대적으로 더디게 나타났습니다. 그러나 두 국가 모두 의료 기술의 발전과 보건 인프라 확장을 통해 사망률은 선진국과 다를게 없이 유지되는 모습을 확인할 수 있습니다. 개발도상국들도 의료 기술의 발전과 보건 인프라 확장을 통해 사망률을 유지 시켜가고 있는 모습으로, 예시로, 방글라데시는 예방 접종 프로그램과 보건 교육을 통해 생명 유지율을 높였고, 나이지리아 역시 의료 인프라 개선을 위한 다양한 노력들이 진행되고 있습니다.

5-4) 머신 러닝 모델

선형 회귀 모델을 구현하여 미래 인구 추세를 예측했습니다.

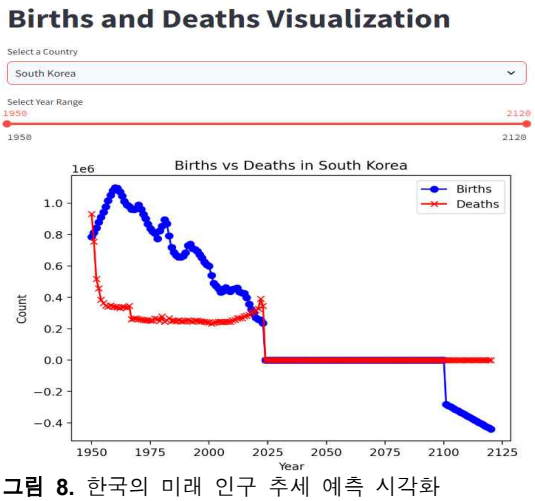

그림 8. 한국의 미래 인구 추세 예측 시각화
Fig. 8. Visualization of predicted future population trends in Korea

위 그림은 한국의 미래 인구 추세 예측을

시각화한 그림입니다. 선형 회귀 모델을 사용하여 미래 출생률을 예측했습니다. 이 모델은 과거 출생률과 연도 간의 관계를 기반으로 미래의 출생률을 예측하는 방식으로 작동합니다. 예측된 값은 2024년부터 2100년까지의 미래 출생률을 포함하며, 이를 통해 출생률의 장기적인 추세를 시각화할 수 있었습니다. 모델의 R^2 점수는 0.92로, 이는 선형 회귀 모델이 과거 데이터를 매우 잘 설명하고 있다는 것을 의미합니다. 이를 바탕으로, 미래의 출생률이 일정한 속도로 감소할 것이라는 추세를 예측할 수 있었습니다.

VI. 결 론

이 연구는 선진국과 개발도상국의 인구 추세 분석을 통해 미래 인구 예측에 대한 중요한 통찰을 제공합니다. EDA 및 선형 회귀 모델을 활용하여 출생률과 사망률의 추세를 분석하고 예측하는 과정을 수행했으며, 이를 통해 선진국에서는 출산율의 감소와 안정적인 사망률을, 개발도상국에서는 여전히 높은 출산율과 사망률 안정을 확인할 수 있었습니다. 특히 한국, 일본, 중국 등의 선진국은 1970년대 이후 출산율 감소 추세를 보였으며, 이에 따라 장기적인 인구 감소가 예상됩니다[5]. 반면, 나이지리아, 방글라데시와 같은 개발도상국은 출산율 감소가 더디게 나타나고 있으며, 사망률 감소가 급격하게 진행되고 있습니다[6]. 미래 예측을 통해 우리는 각 국가의 인구 구조 변화에 대한 심층적인 이해를 얻을 수 있었고, 선형 회귀 모델을 기반으로 미래 출생률을 예측하여 장기적인 인구 변화에 대한 중요한 통찰을 제공했습니다. 이러한 예측 결과는 정책 입안자들이 인구 정책을 수립하는 데 중요한 기초 자료로 활용될 수 있습니다[7]. 향후 연구에서는 이주 통계, 기후 변화 지표 등 추가적인 변수들을 고려하여, 인구 역학에 대한 보다 포괄적인 모델을 구축하고, 기후 변화나 사회적 요인이 인구 변화에 미치는 영향을 분석하는 방향으로 나아갈 것입니다[8]. 이러한 연구는 글로벌 인구 문제를 해결하기 위한 중요한 정책적 방향을 제시할 수 있을 것입니다.

References

[1] United Nations, "World Population Prospects 2022".
[2] Kaggle, "Population by Country 1950-2100", Retrieved from [https://www.kaggle.com]
[3] G. James et al., "An Introduction to Statistical Learning", Springer, 2013.
[4] .J. Tukey, "Exploratory Data Analysis", Addison-Wesley, 1977.
[5] Author, A., & Author, B. (Year). Title of the paper or report. Journal Name, Volume (Issue), page range.
[6] Author, C., & Author, D. (Year). Title of the book. Publisher.
[7] Author, E., & Author, F. (Year). Title of the article. Website. URL
[8] Author, G., & Author, H. (Year). Title of the report. Organization Name. URL