Analysis of Contemporary Human Lifespan

by Sukeerth Kalluraya, Malvika Manohar, Behzod Mirpochoev, and Suresh Nayak - Group 15

Description

The Age dataset contains structured information on the life, work, and death of more than a million deceased famous people. The people in this dataset belong to a variety of groups defined by nationality, gender, occupation, era, etc. Our project focuses on mining this data set to extract any meaningful relationships between these groups. In particular, we plan to understand how the general life expectancy of a person is impacted by the attributes contained in the data set. Some of the interesting questions that we're hoping to answer are:

- Are famous people of a certain region more likely to die young than others?
- Among famous personalities, is the percentage of death by suicide more skewed towards either gender?
- Which occupations involve a higher degree of accidents leading to death?
- Has the general life expectancy of famous people increased with time?

Prior Work

Paper	Author	Work	Data	Methods & Tools
Age dataset: A structured general-purpose dataset on life, work, and death of 1.22 million distinguished people	Annamoradnejad, Issa; Annamoradnejad, Rahimberdi (2022)	Primary motivation was age collection, with secondary emphasis on human descriptor subcategories	Age, Nationality, Post, Gender, Era	Regex, XGBoost classifier, SVM classifier

Prior Work

Paper	Author	Work	Data	Methods & Tools
Public health utility of cause of death data: applying empirical algorithms to improve data quality	Johnson, S.C., Cunningham, M., Dippenaar, I.N. et al. (2021)	Attempting to specify codes that are not specific enough in cause of death	Age, Country of Origin, Cause of Death	Negative correlation, proportional redistribution

Prior Work

Paper	Author	Work	Data	Methods & Tools
Comparison of machine learning algorithms applied to symptoms to determine infectious causes of death in children	Susan Idicula-Thomas, Ulka Gawde and Prabhat Jha (2021)	Comparing techniques of classification to determine cause of death	Age, Cause of Death, Disease Code, ICD-10 Codes, Number of Cases, % Cases	SVM, GBM, CS, ANN, kNN, CART

Prior Work

Paper	Author	Work	Data	Methods & Tools
Estimating expected life-years and risk factor associations with mortality in Finland	Härkänen, Tommi, et. al. (2020)	Estimating the life expectancy of an individual by breaking down risk factors into certain categories to determine most vital risk factors that play in age	Age, Socio-economic, Medical History, Lifestyle, Life Rating, Biological Risk Factors	Poisson Regression Model, Multiple Imputation, Simple & Multiple Regression Models

Prior Work

Paper	Author	Work	Data	Methods & Tools
Comparative Analysis of Suicide, Accidental, and Undetermined Cause of Death Classification	Gray, Douglas, et. al. (2014)	Importance placed on risk factors associated with low frequency events causing death and how to classify these events as such	Age, Method of Death, Injury, History of Ideation, Suspicion	SAS, Logistic Regression, Multinomial Logistic Regression

Datasets

Data Set: https://www.kaggle.com/datasets/imoore/age-dataset

- The data has been provided by the ICWSM.
- This structured dataset is available as a csv file.
- There are 1,222,997 data objects where each data object represents a historical figure.
- The dataset contains 10 attributes for each data object including ID, Name of the notable person, Short description of said person, Gender, Country, Occupation, Birth year, Death year, Manner of Death and Age of death.

	Id	Name	Short description	Gender	Country	Occupation	Birth year	Death year	Manner of death	Age of death
0	Q23	George Washington	1st president of the United States (1732–1799)	Male	United States of America; Kingdom of Great Britain	Politician	1732	1799.0	natural causes	67.0
1	Q42	Douglas Adams	English writer and humorist	Male	United Kingdom	Artist	1952	2001.0	natural causes	49.0
2	Q91	Abraham Lincoln	16th president of the United States (1809-1865)	Male	United States of America	Politician	1809	1865.0	homicide	56.0

All team members have the dataset downloaded.

Proposed Work

(Data Cleaning and Data Preprocessing)

- Handle missing data in attributes such as Death year, Gender, Short Description, etc.
- The attribute Death year has only one missing value; for the creator of this dataset. Remove this data object.
- Convert the values of the gender attribute to consist forms. Eg: “Transgender Male; Male” to “Transgender Male” and “Female; Female” to “Female”.

Proposed Work

(Data Cleaning and Data Preprocessing)

- The country attribute contains values of the country's name in the present era and also in some cases it includes the empire it was ruled by decades ago. Convert the values of the Country attribute to consist forms (i.e., the country's current official name).
- Handle outliers, including data fields that have Death year prior to 1700 and those that have multiple occupations listed under the Occupation attribute.
- Encode the values of categorical attributes such as Country names to numerical values for the purpose of feeding the attribute to ML models using techniques such as One-hot encoding, Label encoding, etc.

Methods & Tools

Tools:

- Python 3
  - pandas: data cleaning and analysis
  - numpy: data transformation and analysis
  - sklearn: procure data adjustment, test-train split, data analysis

Methods:

- Perform Descriptive analysis by plotting and charting each attribute to better understand and summarize the patterns already present in the data; this would also aid in further feature engineering
- We intend to perform data analysis techniques we deem fit after comparison of such techniques against one another with empirical data
  - Possible classification / clustering methods:
    - kNN
    - Decision Trees
    - SVM
    - etc.

Evaluation

As aforementioned, we intend to compare our classifications of life expectancy based on risk factors, cause of death, and group relationships through comparison with the empirical data. We intend to verify our model through performance measures, such as:

- Accuracy
- Precision
- Specificity
- Recall
- etc.