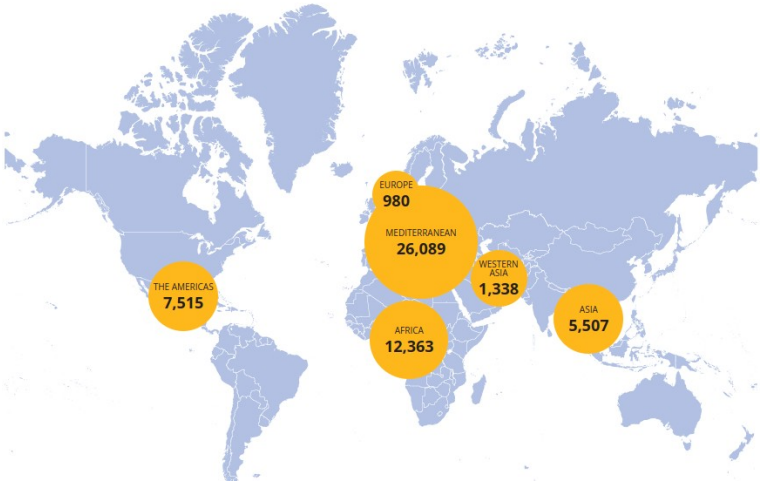
Migration Death Rate Prediction

Zhiyun (David) Lu

Introduction

- Motivation:
  - Displacement has become a reality for many
  - This process can often be deadly, and intentionally made more deadly by institutions

DEATHS DURING MIGRATION
RECORDED SINCE 2014, BY REGION OF INCIDENT



Introduction

- Question: What is the death rate of migrants in a deadly incident given related information?
- Dataset: Missing Migrants Project
  - https://missingmigrants.iom.int/data

Dataset Introduction

- Scope: Migrants (No matter legal status) who have died at the external borders of states, or in the process of migration since 2014
- List of features: Main ID, Incident ID, Incident Type, Region of Incident, Website Date, Incident year, Reported Month, Number of Dead, Minimum Estimated Number of Missing, Total Number of Dead and Missing, Number of Survivors, Number of Females, Number of Males, Number of Children, Region of Origin, Cause of Death, Country of Origin, Migration route, Location of death, Information Source, Coordinates, UNSD Geographical Grouping, Article title, Source Quality, URL
    - Don’t worry, we will cover what they mean when we use the feature
    - Documented on https://missingmigrants.iom.int/methodology but outdated

Dataset Limitations - Background

- The website was created and maintained with the financial support of the governments of Germany, Switzerland and the United Kingdom.
- Data on migrant fatality is inherently hard to collect
  - Bodies are hard to find quickly
  - Lack of consistent reporting
  - Over-represents parts of the world with better media coverage

Dataset Limitations - What is Excluded

- Doesn’t contain information about the general migrant flow
  - We won’t be able to put the death count into perspective
- Doesn’t include death after arrival
  - Immigration detention facilities
  - deportation to a migrant’s homeland
  - labor exploitation
- Doesn’t include internal displacement
- See https://missingmigrants.iom.int/methodology for limitations

Preliminary Cleaning - Column Removal

- Remove features we definitely don’t need
- IDs
  - Main ID
  - Incident ID
- Features Attributing Specific Source
  - Information Source
  - Article title
  - URL

```
#drop columns we definitely don't use
data = raw.drop(['Main ID','Incident ID','Information Source','Article title','URL'],axis=1)
```

Preliminary Cleaning - Row Removal

- Incident Type
  - “Incident” “Cumulative Incident” “Split Incident”
  - We only want “Incident” for single incidents
- Number of Survivors
  - Not always available
  - We need it for calculating the death rate

Incident Type

- Code

```
# include only entries for individual incidents
data.drop(data[data['Incident Type']!='Incident'].index,inplace=True)
data.drop(['Incident Type'],axis=1,inplace=True)
```

- Effect
  - Go from 12398 entries to 12072 entries

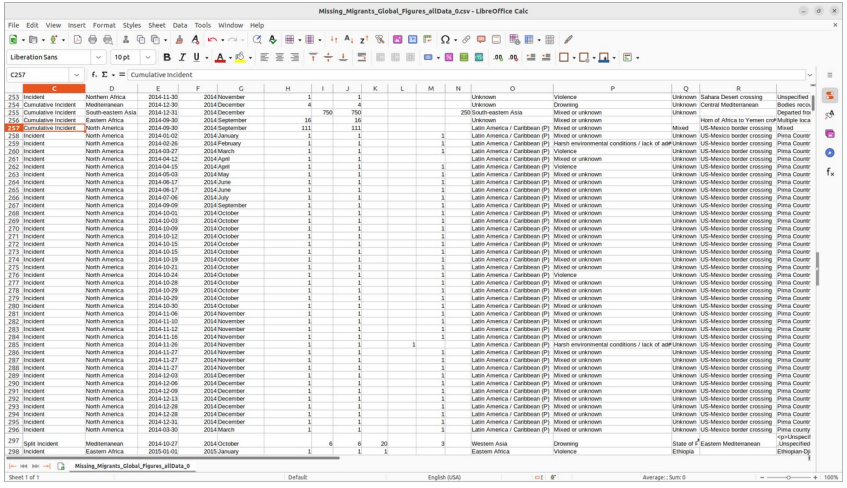
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12072 entries, 0 to 12397
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Region of Incident                    12072 non-null  object
 1   Website Date                          12057 non-null  object
 2   Incident year                         12072 non-null  int64
 3   Reported Month                        12072 non-null  object
 4   Number of Dead                        11612 non-null  float64
 5   Minimum Estimated Number of Missing   1072 non-null   float64
 6   Total Number of Dead and Missing      12072 non-null  int64
 7   Number of Survivors                   1734 non-null   float64
 8   Number of Females                     2145 non-null   float64
 9   Number of Males                       7466 non-null   float64
 10  Number of Children                    1418 non-null   float64
 11  Region of Origin                      12071 non-null  object
 12  Cause of Death                        12072 non-null  object
 13  Country of Origin                     12065 non-null  object
 14  Migration route                       9198 non-null   object
 15  Location of death                     12072 non-null  object
 16  Coordinates                           12036 non-null  object
 17  UNSD Geographical Grouping            12071 non-null  object
 18  Source Quality                        12071 non-null  float64
dtypes: float64(7), int64(2), object(10)
memory usage: 1.8+ MB
```

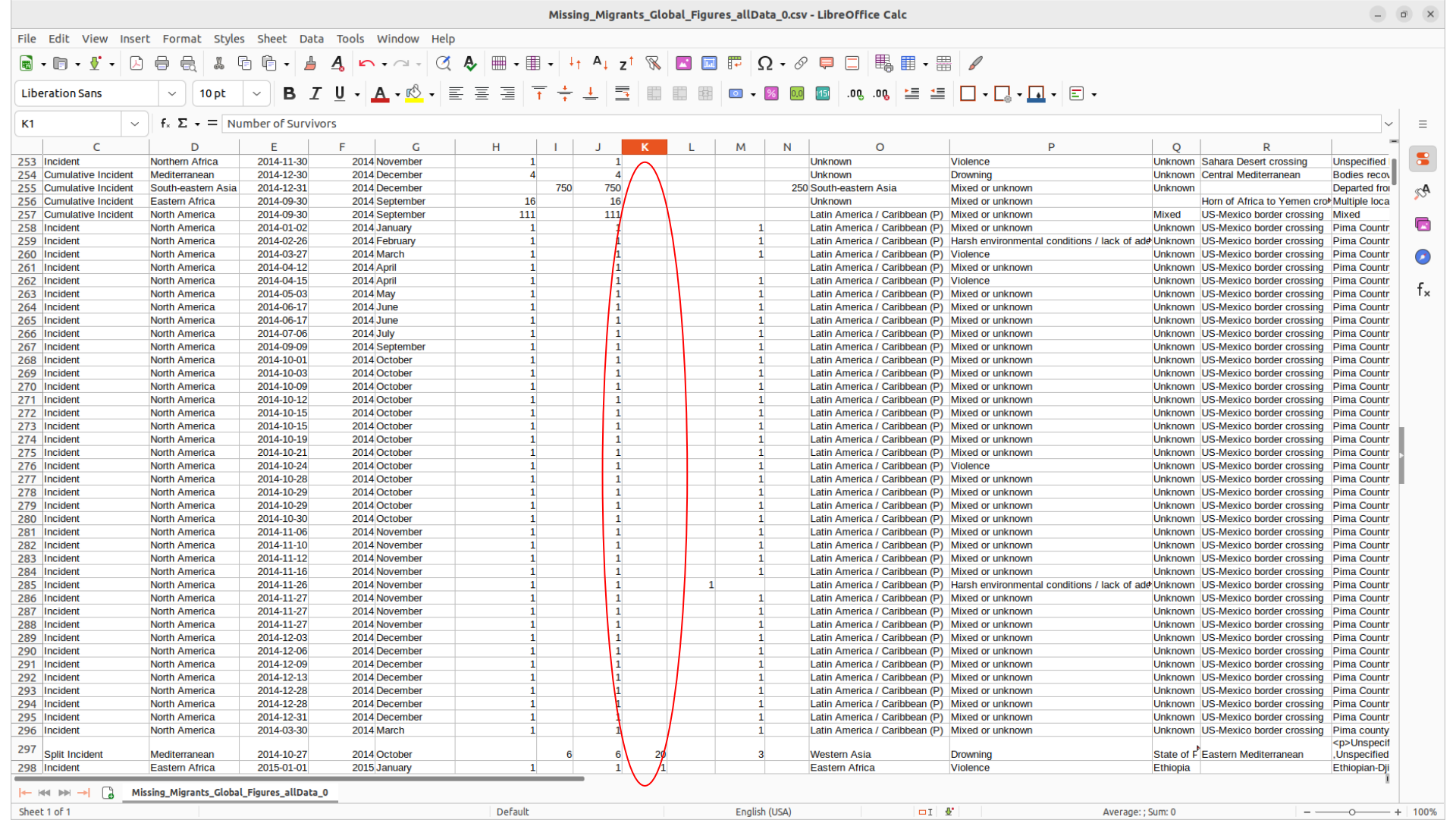
Number of Survivors

- Absent from a majority of rows

```
7   Number of Survivors          1734 non-null   float64
```

  - Constant imputation isn't a good idea
- Systematically missing from certain sub-categories of our data
  - KNN-imputation is not a good idea
  - Semi-supervised learning might not produce the best results
  - Skewing of composition caused by removing rows missing this feature





	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
253	Incident	Northern Africa	2014-11-30	2014	November	1		1					Unknown	Violence	Unknown	Sahara Desert crossing	Unspecified
254	Cumulative Incident	Mediterranean	2014-12-30	2014	December	4		4					Unknown	Drowning	Unknown	Central Mediterranean	Bodies recov
255	Cumulative Incident	South-eastern Asia	2014-12-31	2014	December		750	750				250	South-eastern Asia	Mixed or unknown	Unknown		Departed fro
256	Cumulative Incident	Eastern Africa	2014-09-30	2014	September	16		16					Unknown	Mixed or unknown		Horn of Africa to Yemen cro	Multiple loca
257	Cumulative Incident	North America	2014-09-30	2014	September	111		111					Latin America / Caribbean (P)	Mixed or unknown	Mixed	US-Mexico border crossing	Mixed
258	Incident	North America	2014-01-02	2014	January	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
259	Incident	North America	2014-02-26	2014	February	1		1			1		Latin America / Caribbean (P)	Harsh environmental conditions / lack of ade	Unknown	US-Mexico border crossing	Pima Countr
260	Incident	North America	2014-03-27	2014	March	1		1			1		Latin America / Caribbean (P)	Violence	Unknown	US-Mexico border crossing	Pima Countr
261	Incident	North America	2014-04-12	2014	April	1		1					Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
262	Incident	North America	2014-04-15	2014	April	1		1			1		Latin America / Caribbean (P)	Violence	Unknown	US-Mexico border crossing	Pima Countr
263	Incident	North America	2014-05-03	2014	May	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
264	Incident	North America	2014-06-17	2014	June	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
265	Incident	North America	2014-06-17	2014	June	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
266	Incident	North America	2014-07-06	2014	July	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
267	Incident	North America	2014-09-09	2014	September	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
268	Incident	North America	2014-10-01	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
269	Incident	North America	2014-10-03	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
270	Incident	North America	2014-10-09	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
271	Incident	North America	2014-10-12	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
272	Incident	North America	2014-10-15	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
273	Incident	North America	2014-10-15	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
274	Incident	North America	2014-10-19	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
275	Incident	North America	2014-10-21	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
276	Incident	North America	2014-10-24	2014	October	1		1			1		Latin America / Caribbean (P)	Violence	Unknown	US-Mexico border crossing	Pima Countr
277	Incident	North America	2014-10-28	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
278	Incident	North America	2014-10-29	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
279	Incident	North America	2014-10-29	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
280	Incident	North America	2014-10-30	2014	October	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
281	Incident	North America	2014-11-06	2014	November	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
282	Incident	North America	2014-11-10	2014	November	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
283	Incident	North America	2014-11-12	2014	November	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
284	Incident	North America	2014-11-16	2014	November	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
285	Incident	North America	2014-11-26	2014	November	1		1		1			Latin America / Caribbean (P)	Harsh environmental conditions / lack of ade	Unknown	US-Mexico border crossing	Pima Countr
286	Incident	North America	2014-11-27	2014	November	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
287	Incident	North America	2014-11-27	2014	November	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
288	Incident	North America	2014-11-27	2014	November	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
289	Incident	North America	2014-12-03	2014	December	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
290	Incident	North America	2014-12-06	2014	December	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
291	Incident	North America	2014-12-09	2014	December	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
292	Incident	North America	2014-12-13	2014	December	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
293	Incident	North America	2014-12-28	2014	December	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
294	Incident	North America	2014-12-28	2014	December	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
295	Incident	North America	2014-12-31	2014	December	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima Countr
296	Incident	North America	2014-03-30	2014	March	1		1			1		Latin America / Caribbean (P)	Mixed or unknown	Unknown	US-Mexico border crossing	Pima county
297	Split Incident	Mediterranean	2014-10-27	2014	October		6	6	20		3		Western Asia	Drowning	State of F	Eastern Mediterranean	<p>Unspecif ,Unspecified
298	Incident	Eastern Africa	2015-01-01	2015	January	1		1	1				Eastern Africa	Violence	Ethiopia		Ethiopian-Dj

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1734 entries, 4 to 12396
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Region of Incident                    1734 non-null   object
 1   Website Date                          1734 non-null   object
 2   Incident year                         1734 non-null   int64
 3   Reported Month                        1734 non-null   object
 4   Number of Dead                        1485 non-null   float64
 5   Minimum Estimated Number of Missing   693 non-null    float64
 6   Total Number of Dead and Missing      1734 non-null   int64
 7   Number of Survivors                   1734 non-null   float64
 8   Number of Females                     456 non-null    float64
 9   Number of Males                       760 non-null    float64
 10  Number of Children                    379 non-null    float64
 11  Region of Origin                      1733 non-null   object
 12  Cause of Death                        1734 non-null   object
 13  Country of Origin                     1734 non-null   object
 14  Migration route                       1364 non-null   object
 15  Location of death                     1734 non-null   object
 16  Coordinates                           1733 non-null   object
 17  UNSD Geographical Grouping            1734 non-null   object
 18  Source Quality                        1733 non-null   float64
dtypes: float64(7), int64(2), object(10)
memory usage: 270.9+ KB
```

```python
# include only entries with a survivor count
data.dropna(subset=['Number of Survivors'], inplace=True)
```

Further removal of features

- We do not wish to further cut down on number of entries
  - Remove “Source Quality”
- We will use “Total Number of Dead and Missing” as the only measure of death
  - Insufficient data on the population composition of the deceased
  - Missing migrants are presumed dead by definition

Handling Location related features

- Features related to the location of the incident
  - Location of death
  - Region of incident
  - Coordinates
  - UNSD Geographical Grouping
- Features related to where the migrants come from
  - Country of origin
  - Region of origin
- Migration route

Location of death

- Supposed to be a descriptive text
- In the original dataset appear almost categorical
- After shrinking becomes way too varied to encode
- Exact location already covered by coordinates

Location of death
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
near Douglas, Arizona, USA
Border between Russia and Estonia
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Waters near Greece while being towed back to Turkey
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
France - Calais
Off the coast of Fort Lauderdale, Florida
Hat Yai hospital, 182 Ratthakan, Hat Yai, Hat Yai District, Songkhla 90110
Ceuta waters
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Italy waters
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
California-Mexico border near San Diego, California, USA
Dominican Republic
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
India
Gulf of Aden (left from Djibouti)
Libyan Coast
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
France - Calais (Killed on motorway)
Bypass Road, Moo 7, Bang Phra subdistrict, Sri Racha district, Chonburi province, Thailand
Moo 13, 38th km of the 4006 Langsuan-Pato Road, Pato district, Chumphon province, Thailand
Rio Bravo, Ciudad Juarez, Chihuahua, Mexico
Gulf of Aden
France - Calais (Hit by lorry in port)
King Ranch, Brooks County, Texas, USA
France - Calais (Body found on dock)
France - Calais (Body found crushed in transporter)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Mediterranean
Waters near Lesvos, Greece
Lake Albert, Uganda/DRC

Location of death
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
near Douglas, Arizona, USA
Border between Russia and Estonia
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Waters near Greece while being towed back to Turkey
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
France - Calais
Off the coast of Fort Lauderdale, Florida
Hat Yai hospital, 182 Ratthakan, Hat Yai, Hat Yai District, Songkhla 90110
Ceuta waters
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Italy waters
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
California-Mexico border near San Diego, California, USA
Dominican Republic
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
India
Gulf of Aden (left from Djibouti)
Libyan Coast
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
France - Calais (Killed on motorway)
Bypass Road, Moo 7, Bang Phra subdistrict, Sri Racha district, Chonburi province, Thailand
Moo 13, 38th km of the 4006 Langsuan-Pato Road, Pato district, Chumphon province, Thailand
Rio Bravo, Ciudad Juarez, Chihuahua, Mexico
Gulf of Aden
France - Calais (Hit by lorry in port)
King Ranch, Brooks County, Texas, USA
France - Calais (Body found on dock)
France - Calais (Body found crushed in transporter)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Pima Country Office of the Medical Examiner jurisdiction, Arizona, USA (see coordinates for exact location)
Mediterranean
Waters near Lesvos, Greece
Lake Albert, Uganda/DRC

Location of death
Border between Russia and Estonia
France - Calais
Ceuta waters
Italy waters
California-Mexico border near San Diego, California, USA
Dominican Republic
India
Gulf of Aden (left from Djibouti)
France - Calais (Killed on motorway)
Bypass Road, Moo 7, Bang Phra subdistrict, Sri Racha district, Chonburi province, Thailand
Moo 13, 38th km of the 4006 Langsuan-Pato Road, Pato district, Chumphon province, Thailand
Gulf of Aden
France - Calais (Hit by lorry in port)
France - Calais (Body found on dock)
France - Calais (Body found crushed in transporter)
Waters near Lesvos, Greece
Lake Albert, Uganda/DRC
Limpopo River, Zimbabwe
Italy waters
Türkiye waters - en route to Greece
On la Bestia, near Arriaga, Chiapas, Mexico
344 Road, 39-40th km, Nong Sue Chang subdistrict, Nong Yai district, Chonburi province, Thailand
Coast of Libya
Dominican Republic, near Boca Chica and Juan Dolio
Coast of Libya
Waters near Greece
Coast of Libya
Petchkasem road, Tamtua subdistrict, Takua Pa district, Phang Nga province, Thailand
International waters between Libya and Italy
Gulf of Aden
Italian waters
Between Libya and Sicily
Highway 304, Chachoengsao-Kabinburi Rd, Muangmai subdistrict, Ratchasan District, Chachoengsao Province, Thailand
Central Med Route - Sicily
Moo 4, Nong Yai subdistrict, Nong Yai district, Chonburi province, Thailand
Morocco/Spain border near Melilla
Coast of Malaysia - Strait of Malacca
TBC - boat went missing, not recovered
Unspecified location between North Africa and Italy. Survivors taken to Pozzallo
Recued off Sicily; taken to Catania, Sicily
Aegean Sea
40 miles off the coast of Libya
Off the southwest coast of Malaysia
Just off the coast of Garabulli, Libya
30 miles off the coast of Tripoli, Libya
off the coast of Kafr el-Sheikh governorate, Egypt

UNSD Geographical Grouping

- Region based on United Nations Statistics Division (UNSD) geoscheme
- Comparing to “Region of Incident”
  - Upside: more detailed grouping of Europe
  - Downside: lumps all locations at sea into uncategorized
  - See https://missingmigrants.iom.int/regional-classification

Region of Incident	Coordinates	UNSD Geographical Grouping
Europe	POINT (28 59.1551)	Northern Europe
Europe	POINT (2 50.9355)	Western Europe
Mediterranean	POINT (-5 35.8793)	Uncategorized
Mediterranean	POINT (16 34.5645)	Uncategorized
North America	POINT (-117 32.5543)	Northern America
Caribbean	POINT (-68 18.4505)	Caribbean
Southern Asia	POINT (89 25.1841)	Southern Asia
Eastern Africa	POINT (39 20.2802)	Uncategorized
Europe	POINT (2 50.9365)	Western Europe
South-eastern Asia	POINT (101 13.2353642)	South-eastern Asia
South-eastern Asia	POINT (99 9.96833)	South-eastern Asia
Eastern Africa	POINT (48 12.5331)	Uncategorized
Europe	POINT (2 50.9686)	Western Europe
Europe	POINT (2 50.950273)	Western Europe
Europe	POINT (2 50.9688)	Western Europe
Mediterranean	POINT (26 39.398814)	Uncategorized
Eastern Africa	POINT (31 1.63236)	Eastern Africa
Eastern Africa	POINT (32 -23.8148)	Eastern Africa
Mediterranean	POINT (14 35.1092)	Uncategorized
Mediterranean	POINT (27 38.669208)	Uncategorized
Central America	POINT (-94 16.242196)	Central America
South-eastern Asia	POINT (101 13.3493133)	South-eastern Asia
Mediterranean	POINT (14 34.3706)	Uncategorized
Caribbean	POINT (-70 18.3079)	Caribbean
Mediterranean	POINT (14 33.1045)	Uncategorized
Mediterranean	POINT (27 37.848433)	Uncategorized
Mediterranean	POINT (16 32.6801)	Uncategorized
South-eastern Asia	POINT (98 8.855529)	South-eastern Asia
Mediterranean	POINT (16 34.7086)	Uncategorized
Eastern Africa	POINT (48 12.5331)	Uncategorized
Mediterranean	POINT (14 36.4935)	Uncategorized
Mediterranean	POINT (15 34.2775)	Uncategorized
South-eastern Asia	POINT (101 13.6702361)	South-eastern Asia
Mediterranean	POINT (12 35.9061)	Uncategorized
South-eastern Asia	POINT (101 13.1394249)	South-eastern Asia
Mediterranean	POINT (-3 35.2783)	Uncategorized
South-eastern Asia	POINT (100 4.66259)	South-eastern Asia
Mediterranean	POINT (17 34.6033)	Uncategorized
Mediterranean	POINT (14.688719 36.222841)	Uncategorized
Mediterranean	POINT (15.8313 36.4317)	Uncategorized
Mediterranean	POINT (26.921962 37.862031)	Uncategorized
Mediterranean	POINT (15 33.6564)	Uncategorized
South-eastern Asia	POINT (115 6.35306)	South-eastern Asia
Mediterranean	POINT (14 32.8343)	Uncategorized
Mediterranean	POINT (13 33.2423)	Uncategorized
Mediterranean	POINT (31 31.5818)	Uncategorized

Coordinates vs Region of Incident

- Coordinates
  - Require parsing
  - Detailed values
  - Available in all but one entry
- Region of Incident
  - Require one-hot encoding
  - Categorical
  - Can be seen as a kind of feature engineering for the coordinates
  - Available for all entries

Anomalies

- One case with missing coordinates
  - Probably lost due to mistake instead of missing from beginning
  - “Location of death” is “Off the coast of Libya”
  - Manually fill in a set of coordinates from an entry with similar location of death description

78	Mediterranean		Uncategorized

- One case with misaligned features
  - Only appear in the table editor
  - Automatically fixed by pd.read_csv()

1537	South-eastern Asia	location of capsize yet to be determined.	POINT (100.7268416 3.228199546)	South-eastern Asia

Parsing Coordinates

```
# Before proceeding to the harder decisions, let's first parse the coordinates

# Off the coast of Libya prepare for the missing entry
# Arguably should be done after test train split, but there really isn't a way to automate this
# Hopefully one entry doesn't change the evaluations too much
default = (16.9684,33.6053)

# function for parsing
def coor_parse(coordinates):
    # NaN is a float
    if not isinstance(coordinates,str):
        return default
    else:
        splitted = coordinates.split()
        lat = splitted[1][1:]
        lng = splitted[2][:-1]
        return float(lat), float(lng)

# apply function
data['lat'] = data['Coordinates'].apply(lambda x: coor_parse(x)[0])
data['lng'] = data['Coordinates'].apply(lambda x: coor_parse(x)[1])

# remove Coordinates
data.drop(['Coordinates'],axis=1,inplace=True)

data.head()
```

Region of Origin vs Country of Origin (Ideal)

- Country of Origin
  - One entry can have multiple countries of origin
  - More varieties (at least 70 or so)
  - More often unknown
- Region of Origin
  - One entry only has one
  - Need to deal with (P) for presumed
  - Fewer varieties for easier one-hot encoding (19)
  - Might be known even if we don’t know the specific countries

Region of Origin - Complications

- Arbitrary use of language
  - “Unknown” is sometimes used interchangeably with “Mixed”
  - In one specific case, “Region of Origin” is even left blank

Unknown	Cameroon,Somalia,Syrian Arab Republic
Unknown	Unknown
Unknown	Unknown
Unknown	Unknown
Southern Asia	Indonesia
Unknown	Unknown
Sub-Saharan Africa (P)	Unknown
Mixed	Egypt,Syrian Arab Republic
Unknown	Unknown
Unknown	Unknown
Sub-Saharan Africa (P)	Unknown
Western Asia	Syrian Arab Republic
Southern Asia	Afghanistan
Sub-Saharan Africa (P)	Unknown
Unknown	Côte d'Ivoire,Eritrea,Guinea,Libya,Mali,Palestinian Territories,Somalia,Sudan,Syrian Arab Republic
Unknown	Unknown
Caribbean	Haiti
Caribbean	Cuba

	Afghanistan,Iraq,Syrian Arab Republic

Potential solutions and why they might not work

- Using the “Country of Origin” to help with KNN imputation
  - Entries with multiple items in “Country of Origin” can mess up KNN
- Change some “Unknown” in “Region of Origin” to “Mixed” based on the number of countries of origin
  - Still not very useful
- KNN imputation without “Country of Origin”
  - Unknown or mixed might mean something

```python
# The missing value is Western / Southern Asia

# Remove (P)
def p_remove(region):
    # NaN is a float
    if not isinstance(region,str):
        return "Western / Southern Asia"
    elif region[-1] == ")":
        return region[:-4]
    else:
        return region

data['Region of Origin'] = data['Region of Origin'].apply(p_remove)
data.drop(['Country of Origin'],axis=1,inplace=True)

data.head()
```

Time related features

- “Website Date” “Incident year” “Reported Month”
  - Just Year Month Day in different forms
- Removing the day
  - For the shrunken Dataset, the complete date act as an ID
  - The day isn’t really meaningful given the time needed to find the body
- Keep the year and month as different features
  - Make it harder for them to be used as identifying IDs
  - Attempt to capture both the yearly recurrent trend and trend by year

Other

- One-hot encoding
  - Need to be done before the split because certain values have become very rare
- “Migration Route” requires KNN imputing
- “Cause of Death” lays into 7 categories and is always available

```
data['Total Number of Victims'] = data['Total Number of Dead and Missing'] + data['Number of Survivors']
data['Death Rate'] = data['Total Number of Dead and Missing'] / data['Total Number of Victims']

data.drop(['Total Number of Dead and Missing','Number of Survivors'],axis=1, inplace=True)

data.head()
```

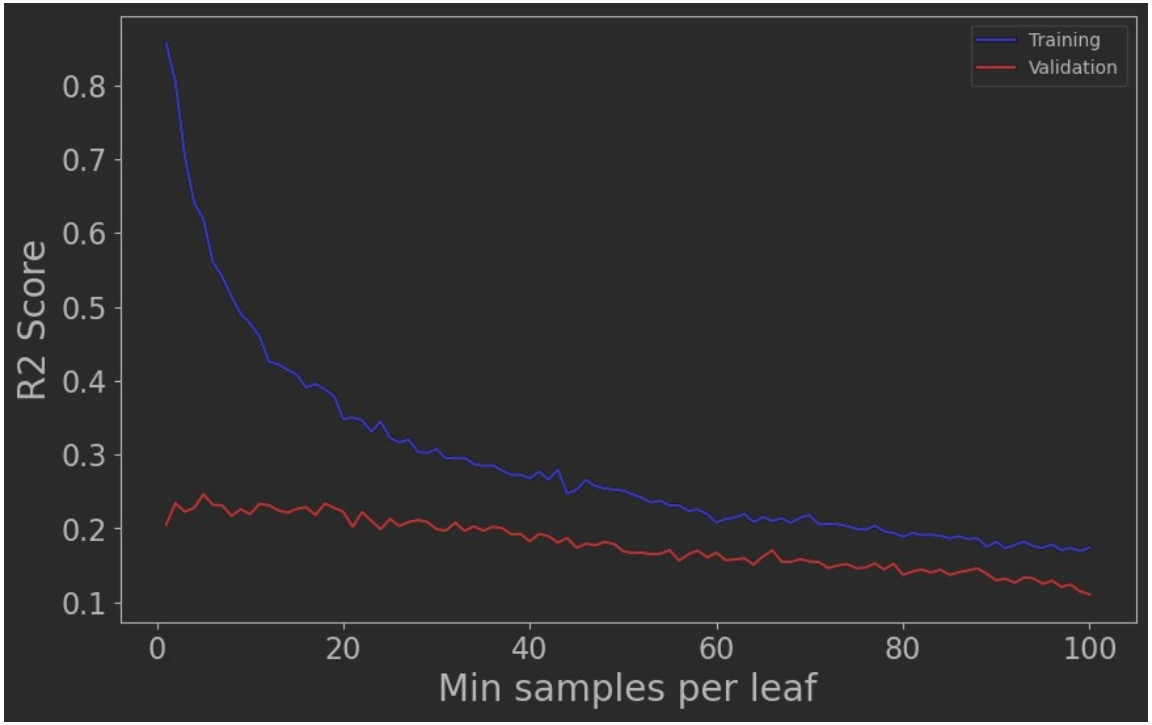
One Unexpected Problem

- Negative number of survivors
  - Cause error by giving infinite death rate
  - Domain knowledge: it is possible, but I lack the proper knowledge to deal with it
  - Go back to make sure the number of survivors is non-negative

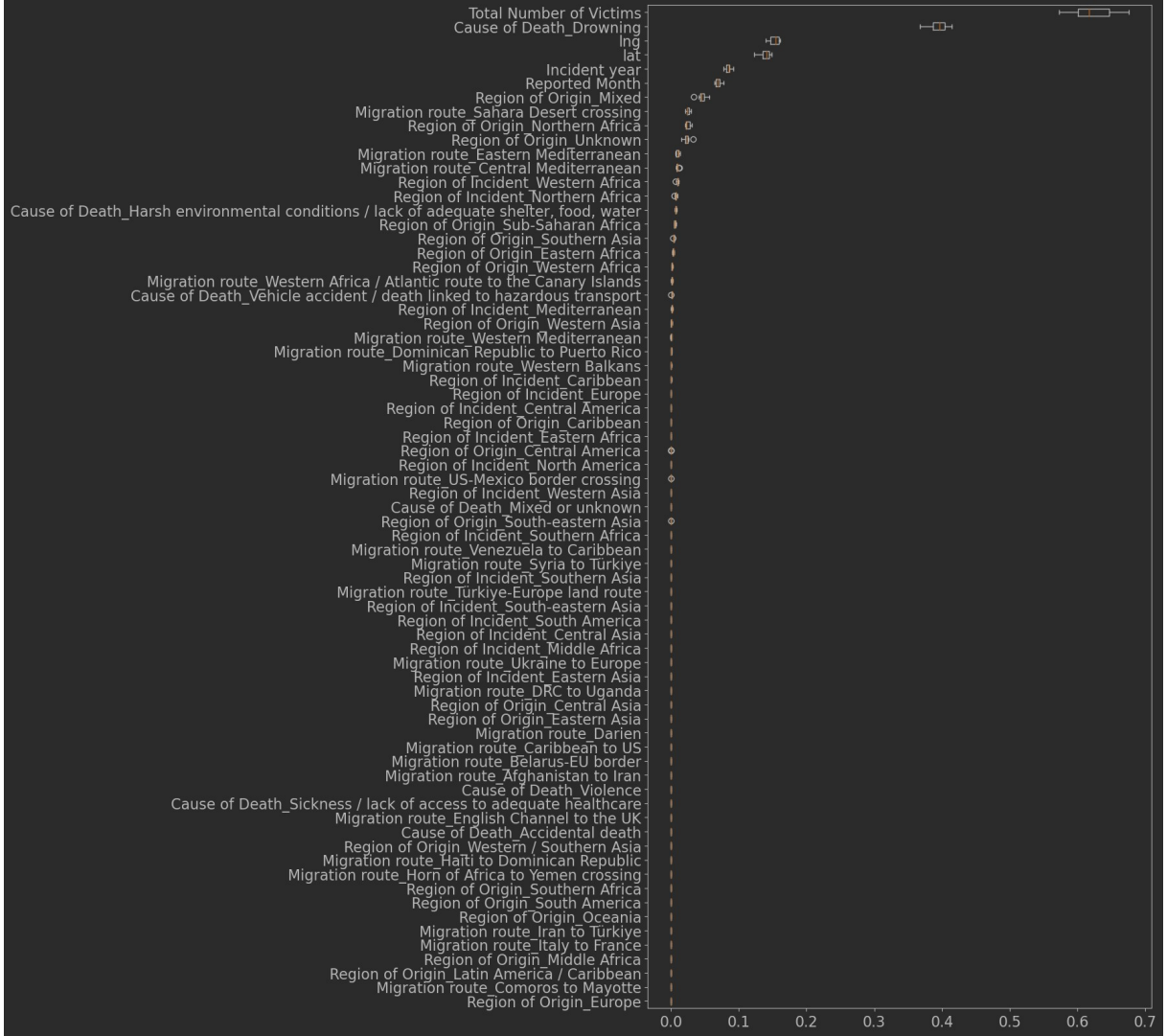
Model of Choice

- Bagging Regressor with DecisionTreeRegressor
  - Doesn’t need standardization
  - Suggested by the Cheat sheet
  - Bootstrapping to deal with somewhat correlated features

Hyperparameter - min samples per leaf

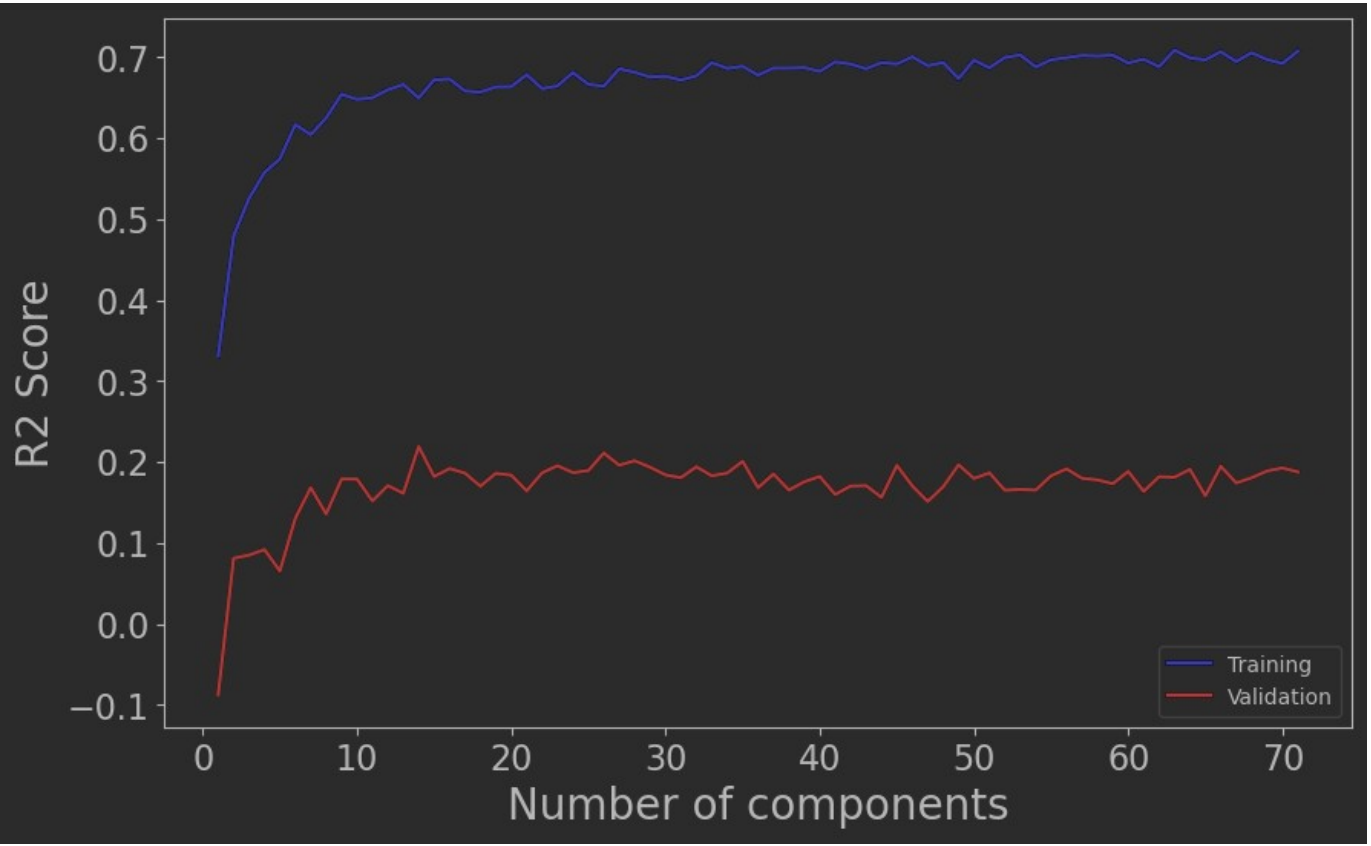



```
Training score:0.61
Testing score:0.30
```


Total Number of Victims
Cause of Death_Drowning
lng
lat
Incident year
Reported Month
Region of Origin_Mixed
Migration route_Sahara Desert crossing
Region of Origin_Northern Africa
Region of Origin_Unknown
Migration route_Eastern Mediterranean
Migration route_Central Mediterranean
Region of Incident_Western Africa
Region of Incident_Northern Africa
Cause of Death_Harsh environmental conditions / lack of adequate shelter, food, water
Region of Origin_Sub-Saharan Africa
Region of Origin_Southern Asia
Region of Origin_Eastern Africa
Region of Origin_Western Africa
Migration route_Western Africa / Atlantic route to the Canary Islands
Cause of Death_Vehicle accident / death linked to hazardous transport
Region of Incident_Mediterranean
Region of Origin_Western Asia
Migration route_Western Mediterranean
Migration route_Dominican Republic to Puerto Rico
Migration route_Western Balkans
Region of Incident_Caribbean
Region of Incident_Europe
Region of Incident_Central America
Region of Origin_Caribbean
Region of Incident_Eastern Africa
Region of Origin_Central America
Region of Incident_North America
Migration route_US-Mexico border crossing
Region of Incident_Western Asia
Cause of Death_Mixed or unknown
Region of Origin_South-eastern Asia
Region of Incident_Southern Africa
Migration route_Venezuela to Caribbean
Migration route_Syria to Türkiye
Region of Incident_Southern Asia
Migration route_Türkiye-Europe land route
Region of Incident_South-eastern Asia
Region of Incident_South America
Region of Incident_Central Asia
Region of Incident_Middle Africa
Migration route_Ukraine to Europe
Region of Incident_Eastern Asia
Migration route_DRC to Uganda
Region of Origin_Central Asia
Region of Origin_Eastern Asia
Migration route_Darien
Migration route_Caribbean to US
Migration route_Belarus-EU border
Migration route_Afghanistan to Iran
Cause of Death_Violence
Cause of Death_Sickness / lack of access to adequate healthcare
Migration route_English Channel to the UK
Cause of Death_Accidental death
Region of Origin_Western / Southern Asia
Migration route_Haiti to Dominican Republic
Migration route_Horn of Africa to Yemen crossing
Region of Origin_Southern Africa
Region of Origin_South America
Region of Origin_Oceania
Migration route_Iran to Türkiye
Migration route_Italy to France
Region of Origin_Middle Africa
Region of Origin_Latin America / Caribbean
Migration route_Comoros to Mayotte
Region of Origin_Europe
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Attempt to reduce features - PCA

- Hold the min samples per leaf still
- Use number of components for PCA as hyperparameter




```
Training score:0.66
Testing score:0.20
```

Conclusion

- Lack of reliable data has made prediction hard
- Permutation Importance
  - Cause of death drowning - overrepresentation of the Mediterranean route
  - Total number of victims - Either travelling in group is safer or it is in fact acting as an ID