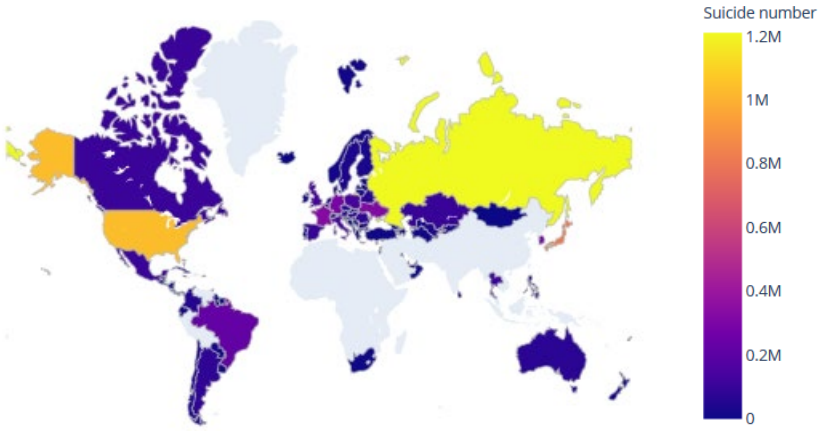
Estimating Suicide Rates Using Data Science

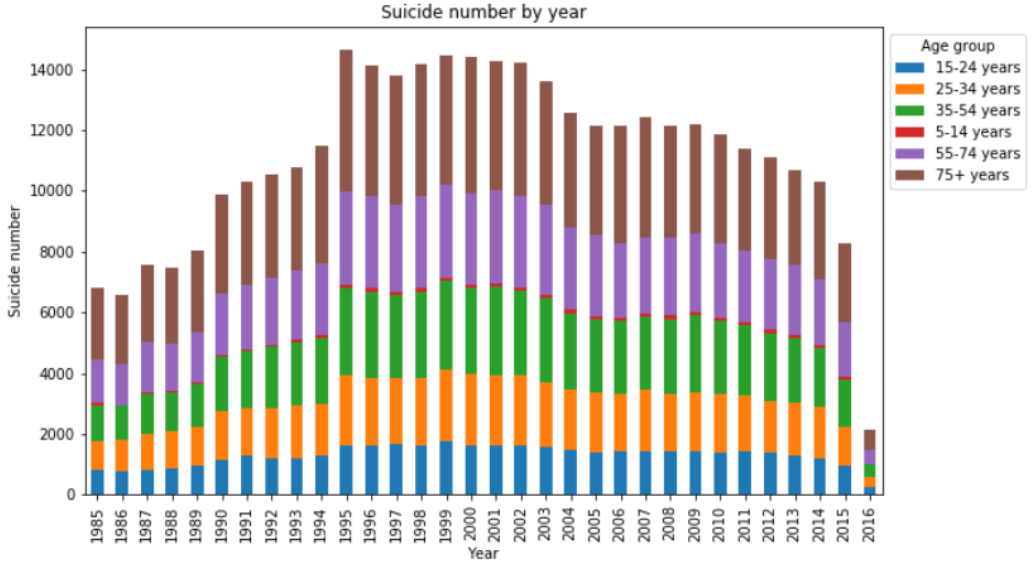
HARI NATH BINGI

Problem

- The cost of suicidal behavior to individuals, families, communities and society makes suicide a serious health issue around the world.
- ABS data (2012) shows more people die from suicide than road deaths.
- Even with various groups offering support there is no significant decrease in the number of suicides globally.
- **What if we can use Data Science techniques to estimate which groups may have high number of suicides. This can help support groups to focus on the right areas.**



Suicides across the world

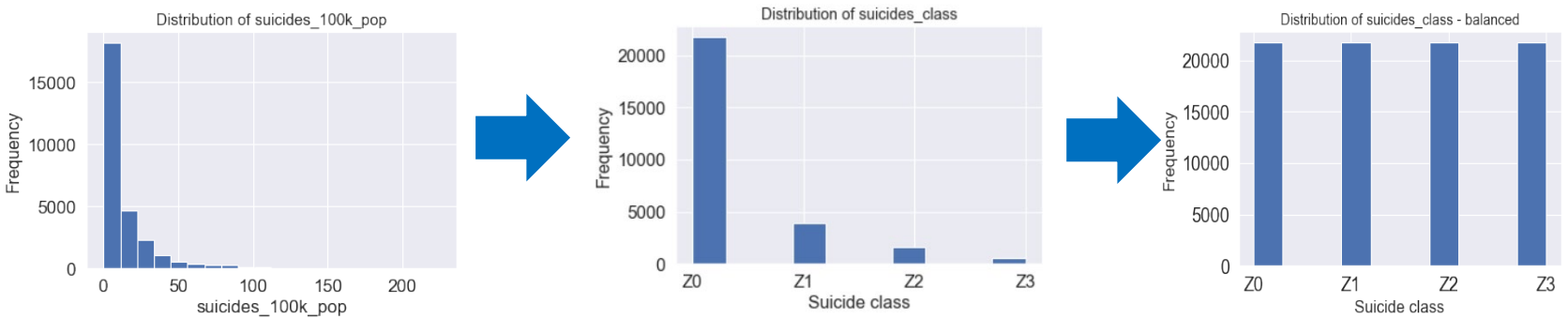


Approach

- Convert numerical variable of suicides to categorical.
- Balance the class distribution.
- Evaluate various classifiers via cross validation using macro f1 as metric.
- One-sided paired t-test to compare classifiers.

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2156624900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2156624900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2156624900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2156624900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2156624900	796	Boomers

Snippet of data set

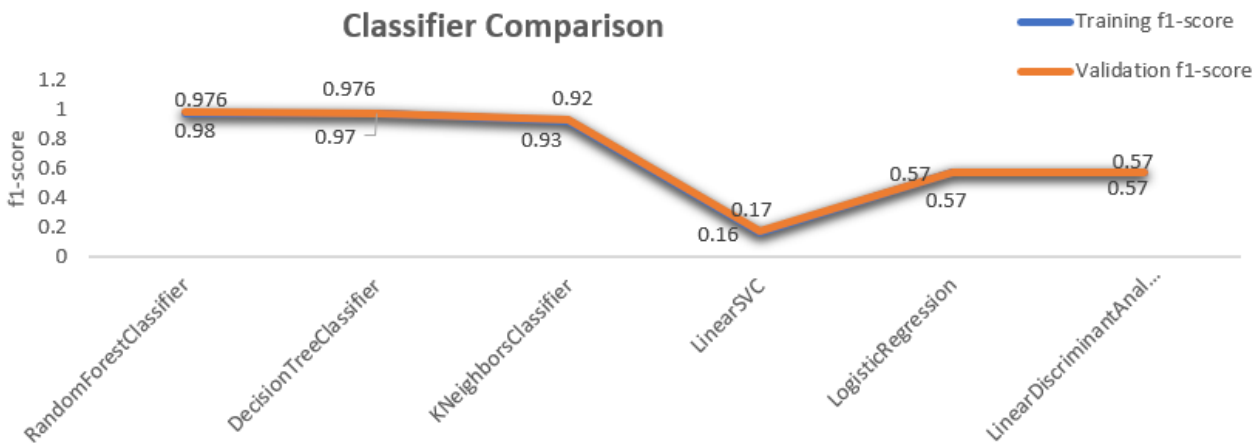


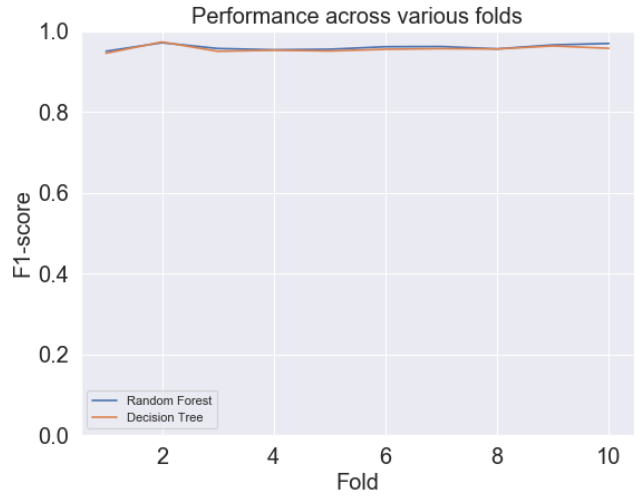
Conversion of suicide count from numerical to categorical and balancing

Class	Value
Z0	0-19 suicide per 100k
Z1	20-39 suicide per 100k
Z2	40-79 suicide per 100k
Z3	80+ suicide per 100k

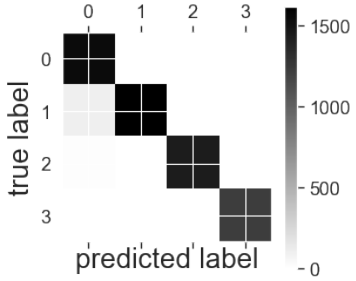
Categorical variable details

Classifier Results





	precision	recall	f1-score	support
Z0	1.00	0.93	0.96	1676
Z1	0.94	1.00	0.97	1626
Z2	0.99	1.00	0.99	1440
Z3	1.00	1.00	1.00	1234
accuracy			0.98	5976
macro avg	0.98	0.98	0.98	5976
weighted avg	0.98	0.98	0.98	5976



Classification Report of Random Forest

- Random Forest Classifier performed the best across various folds of data.
- Random Forest has good performance for every class.
- This indicates the outcome is decision based.
- Linear SVC performed the worst.


THANK
YOU!