SUICIDE RATE PREDICTION USING MACHINE LEARNING

Dissertation submitted in part fulfilment of the requirements
for the degree of Master of Science in Data Analytics
at Dublin Business School

Jobin Lawrence Joseph

10389504

Supervisor: Abhishek Kaushik

MSc in Data Analytics May 2019

Declaration

I, Jobin Lawrence Joseph , declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this work and this work is fully compliant with the Dublin Business School's academic honesty policy.

Signed: Jobin Lawrence Joseph

Date: 17-May-2019

Acknowledgement

It would not have been possible to write this dissertation without the help and support of the benevolent people around me, It is my great privilege to convey my utmost respect and sincere gratitude to Dr. Abhishek Kaushik , my respected supervisor, for his precious guidance and motivation in making this thesis complete. It is sincerely acknowledged that he gave valuable advice all around the entire process and willingness to mentor me without hesitation.

I would also want to convey my heartfelt thanks to Mr. Sreenath , my friend for being there and providing the encouragement during the difficult times and made me optimistic to achieve my goals.

Finally, I use this opportunity to convey my sincere thanks to all of my friends and family who supported me all through my thesis.

Jobin Lawrence Joseph
Dublin, May 2019

Abstract

There is no other place in the universe where life exist ,and the world we live now is sustained by deaths and births , and suicide is being booming around the world as WHO states.In this thesis i am trying to predict the suicide number of countries using five different machine learning algorithms and trying to find which among them is the mosy suitable algorithm for the dataset that i am using. The data set is record of suicides happened around the globe categorized according to different factors such as age , sex, year, country etc.It was found that decision tree was the apt algorithm for the dataset as it ha dthe leasr RMSE score.I think this thesis will be afoundation people who wish to work on suicide prediction in a way that they could help the government and poeple to tackle a upcoming crisis and avoid a hara-kiri situation amomg the communities

Table of Contents

1. Introduction

Suicide is an act of ending one’s own life consciously and it has been derived from a Latin word suicidium which means “to kill oneself” (Gregory, n.d.). It is a self-inflicted death. It is the third main cause for the death in the world. According to World Health Organization, approximately one million people die from suicide every year and based on research, suicide tries increase enormously throughout immature years (Anon., n.d.) Suicide propensity in individuals is a genuine concern which isn't limited to a specific state or nation. The overall insights demonstrate that suicide is a noteworthy supporter of the world's death rate. Also, suicide is one of the 20 essential drivers for death on the planet and represents 5-6% of all passing’s. While suicide is universal, it is prevalent in certain gatherings of individuals. For example, individuals with bipolar scatters display high self-destructive inclinations. It is astonishing to realize that 1 out of 12 teenagers on the planet endeavours suicide. Among the individuals who kick the bucket because of suicides yearly, individuals of age bunch 15 to 29 having the most noteworthy commitment . By 2020, the World Health Organization predicts that passings to this adverse sickness will be over 2% of the worldwide weight of infections (Anon., n.d.)`

It is observed that committing suicide is a big crime and it is a huge loss to the society. It needs a serious attention and proper system to eradicate the plague spreading and save people from doing it by implementing a proper suicide analysis and prevention using clinical data available.

There are mainly three factors which affects an individual in committing suicide,

- Individual factors and experience
    - Physical and emotional problems
    - Puberty and religious beliefs
    - Failure in management strategies
    - Problems in marriage life
    - Failure and level of education ,etc
- Family factors
    - Family structure
    - Family relation and economic factors
    - Health conditions

- Social factors
    - Media influence
    - Unskilled crisis facing

1.2 Background

A growing number of global research have proven that charges of suicide have been related to economic activity. except for Finland, which skilled an boom in suicide rate all through an monetary upswing among 1985 and 1990 and a decline in the course of a subsequent duration of recession, there had been few empirical examples of suicide growing all through instances of monetary prosperity. Despite the fact that research from Eire showed no affiliation between socioeconomic factors and suicide after accounting for time developments, maximum studies have advised that suicide rates generally tend to say no throughout instances of financial prosperity and boom during intervals of recession. In a latest assessment of the populace-stage mental fitness results of monetary downturns, inferred a advantageous association among economic crisis and the onset of psychopathology, inclusive of suicide. It showed a negative association between monthly degrees of financial hobby and fees of suicide in metropolis. but, it did not locate growth in running-aged females (16–64 years) residing in municipalities with ≥10,000 population to be associated with rising unemployment in any of the instances, which indicated that the connection between suicide and economy is complex and warrants in addition investigation (Yin HL, 2008).

Suicide risk prediction have been studies since decade . It has been noticed that countries with low GDP rates per year tends to record the maximum number of suicides per year that the one s which flourishes its economic base .Its not only the economic factors affects the suicidal tendency but the mental and physical well being of a person has to be confirmed.

1.3 Motivation

According to the Centre for Disease Control and Prevention (CDC) ,it’s been said that the third leading death cause for 15-24 yeas old beings i.e. teenagers are suicide. It created a huge impact on me and I got wonderstruck and was curious to know the facts and factors that influence you g adults to commit suicide. As far as I understood most of the suicide happens when a situation or a crisis occurs and people who are facing them are not ready or capable to tackle them or they are unskilled to go through that phase .This will lead them to a heavy depression and finally they end up committing suicide

I thought of analyzing the problems regarding the booming number of suicide that took place in different countries due to socio, economic or individual causes by providing ways to prevent it in the future by analyzing there democratic , economic and social conditions .I wish to help the government by forecasting a crisis and helps them to take measures that decreases the impact of the problems caused by the crisis and helps people to better understands what's coming on the way and helps them to tackle the situation wisely.

1.4 What should we need to achieve?

The main objective in this research is to use optimal algorithm to improve accuracy in predicting and classifying most influential factors that helps in achieving to find out the crucial factors that pokes out the suicidal tendency in different countries. This paper investigates present studies on various ways in which suicide evaluation and predictions have been accomplished. a number of those methods encompass –Social Networking sites (SNS) based totally suicide detection systems using gadget gaining knowledge of and text mining techniques, suicide analysis from census facts in India ,suicide analysis via the facts acquired from various questionnaires, suicidal tendencies detection from diverse blogs and additionally from numerous records sets available on suicide globally. This paper may even look at the shortcomings of present research and also spotlight destiny research to enhance existing work. . Data analysis is executed, in order to classify statistics on the way to offer a fixed of preventive measures to manipulate them in destiny. Data Set on which the analysis could be completed will include parameters listed below, state, year, kind (reason), Gender, Age group which might assist to provide preventive measure for suicide to an character belonging to positive age group, of a specific state or of a selected gender. The records evaluation can provide records approximately the motive of suicide taken in a specific nation observed via in a particular yr. The records Set also can provide statistics approximately whether the Suicide fee for a particular cause has improved or no longer. Observed by using a prediction what can be the Suicide price in destiny? it is able to additionally provide records on what all regions need to be appeared upon to lessen the Suicide price in special nation as a result of extraordinary reasons. Analysis and type will no longer best provide preventive measures however may even brought about assessment as a wayto offer statistics whether suicide charge has extended or reduced in numerous years. Obs erved with the aid of answer to impeach whether or not it is able to growth or lower in the subsequent coming years

1.5 Research Question

Using five of the machine learning algorithms such as

- Linear regression
- Decision Tree
- Random Forest
- Support Vector Machine
- K- Nearest Neighbour

Using the above algorithms we are trying to predict the suicide number, and trying to find out the best algorithm for getting an accurate result.

1.6 Roadmap of thesis

The proposition is carefully dealt with into straight out information that will empower us to accomplish our normal targets. The essential segment shows how suicides are foreseen dependent on explicit factors. It gives a reckless investigate AI and how it separates the purposes behind an individual to end everything. We have portrayed the issue and bounce into the current circumstance on the planet; this leads us to our purpose of foreseeing the suicide rate. , by recognizing key factors that impacts an individual and a whole system or nation.

The accompanying part examines study of existing composition on AI estimations that can be associated with grasp delegate trimming down rates. We start with an illumination of the set up theories and move towards contemporary hypotheses and thoughts. The observational examination slants a gaps and investigation on completing AI estimations to envision suicide rates. We close with a layout of system and the results we envision from the examination

1.7 Scope and Limitation

Given its particular nature, investigation on suicide faces an arrangement of constraints that forbid advance inside the getting, counteractive action, and treatment of the inconvenience. Because of the truth the world will be a combination of assorted orders that grew up severally, issues with understanding base investigation produce issues with correspondence, language, and disciplinary contentions. Additionally, selecting scientists to the globe is inconvenient due to the different snags that the world appearances, as referenced at some phase amid this budgetary ruin. As demonstrated in monetary ruin one, the expression utilized among suicide scientists is conflicting. Thusly, it's way vigorous to incite trustworthy numbers pretty much the pervasiveness and frequency of suicide and suicide attempts. in activity with sufferers that blessing a chance of suicide offers good and issues of security so it'll be strenuous to cure (Sciences., 2002). Particular measures ought to be taken to blast the implemented science intensity of intercession and obstruction investigation, for the goal that suicide can be a sensibly phenomenal occasion. Those procedures differ from abuse trade endpoints like dangerous higher subjective procedure to discovering methodologies to development the elements of the populace. Investigate on suicide is plagued with a couple of system bothers that containment progresses inside the field. Definitions need consistency, proximal measures don't give off an impression of being customarily prophetical of suicide, reportage of suicide isn't right, and its low repeat exacerbates these issues. The base-charge of finished suicide is acceptably low to square nearly the main basic of examination. While such examination is done, resultant relationships are between phenomenally little and giant relationship of individuals (suicide completers as hostile non-suicide completers, or suicide attempters as unpleasant non-suicide attempters). Usage of heedless change as last results will construct normality and alleviate the issue to a limited degree; in any case, it's so much dubious whether futile transformation could be a strong pointer of suicide last touch (Health., 2008). Crafted by each endeavour and fulfilments will perplex the examination on record that attempters may to boot speak to arrangement of the suicides completed inside the research time frame. Because of the evident reality the proportion of the check examination rushes to try and consider amassing agreeable records on the low repeat endpoints of suicide or suicide attempt, proximal measures containing changes in limit or demeanour are used. In any case the prophetical worth of these elements is unconfirmed. Associated math techniques and proximal endpoints may offer answers; in any case a gigantic open base is perfect (Health., 2008).

Most studies have found that the optimal time to conduct this kind of investigation is between 2 and 6 months after the death. Informants' emotions may be too raw to conduct an extensive interview prior to 2 months after the death. Longer than 6 months after the death, many informants want closure on the suicide and no longer are willing to open up and discuss emotionally difficult topics. The quality of information, measured by the number of diagnoses generated, did not vary as a function of the amount of time since the death .Caution should be used when interpreting information gathered from friends and relatives; one experimental study found that subjects' descriptions of psychological distress varied with characteristics of the deceased and aspects of the manner of death. Use of a comparison group of individuals who died accidentally by similar means could strengthen validity of findings. In general, when case-control methods are used in psychological autopsy studies, the comparisons are made to individuals who died by natural causes matched on demographic variables or psychiatric diagnoses (Center, n.d.).

The psychological autopsy has many similarities to the Family History-Research Diagnostic Criteria or any other indirect interview. The interview is less informative than a direct interview but improves with the number of informants. Certain informants may provide specific information that may not be available from others. For example, friends of adolescent suicide victims may be more aware of substance use and abuse than parent’s .Employers and co-workers may be able to describe the victim's functional ability on the job; for younger victims, interview of teachers and review of school records may play an analogous role. Certain types of information are very difficult, or even impossible, to obtain with a psychological autopsy approach. For example, sexual orientation is information that the victim may have been subliminally aware of, or may not have confided to a friend or parent. Information processing style, or other laboratory-based measures obviously cannot be obtained without the victim's self-report. However, psychological autopsy studies can help to identify living individuals whose characteristics closely resemble suicide victims who can then be studied using more dynamic assessments (Anon., 2017; Ribeiro, 2017). Most examinations have affirmed that the chief spellbinding time to lead this kind of examination is among a few and about a half year when the passing. Witnesses' suppositions is moreover too much unrefined to lead accomplice complete gathering before a few months when the loss of life. Longer than about a half year when the dying, a couple of observers need end at the suicide and now not are inclined to open up and talk showing up notable subjects. The amazing of records, evaluated by technique of the contrast of conclusions created, did never again change as a work of the measure of your time for the illumination that passing .advised must be used once coding convictions amassed from amigos and loved ones; one test watch set that focuses' portrayals of mental torment moved with properties of the died and parts of the procedure of death. Use of a refinement establishment of this World Health Organization kicked the pail by chance through for all intents and purposes indistinguishable strategy must sustain authenticity of disclosures (Ribeiro, 2017). At the point when all is said in done, while case-control ways are used in mental after death examination ponders, the connections are made to individuals that went on through flavourer reasons facilitated on measurement components or solution investigations.

2. Literature Review

2.1 Historical Perspective

Suicide was taken into consideration a serious offense in almost all Western EU countries from the Middles ages till (at the least) the French Revolution (Anon., 1997). In England, one of the last EU nations to decriminalize suicide, 'self-homicide' became against the law up until 1961 (Caruso, n.d.) . In many nations, which include Singapore, its miles nevertheless considered against the law nowadays. Because of this, historic statistics on this subject matter isn't easily to be had; indeed, as we talk within the records great section, the stigma surrounding suicide makes dimension tough even nowadays.

Self slaughter is an extreme situation that causes terrible pain and loss among hundreds and thousands of human beings every year. Each suicide is an adversity. According to World Health Organization (WHO) over eighty thousand people die to suicide every year. As we co-relate according to age there are around 115 per hundred thousand people kill themselves , that is a figure equivalent to a person killing themselves every thirty seconds.

According to World Health Organization's (WHO) fatality data , suicide tendency varies from different community in different groups are higher .One important origin of divergence is both globally and within nations are the sex or gender. As per World Health Organization ,suicidal tendency are maximum in male comparing with the females that too in higher economic boosting countries

It is possible to pin down it's cause and it is very complex to analyse and with the ghelp of some computational statistics it can be analysed in some margins. Mental health, depression ,economic background are some vital roles which plays in human life for the suicidal tendency to grow up.

2.2 WHO Suicide Rate

Preventing Suicide: a global imperative – a report published by World Health organization in 2014 targets increasing awareness for the public's health significance o suicide . It also deals with making suicide prevention a high priority on global public health agenda. As per World Health Organization , Russia is the country which accounts the most number of suicide followed by Japan and the United States

2.3 IHME Suicide Rates

Institute of Health Metric and Evaluation standardizes their death ratio from suicide/self harm measured as the number of deaths per hundred thousand people and as per them in 2017, Russia accounts the most suicide and counted at 2509 deaths per hundred thousand individuals . They standardize death according to the age and as per them age standardization

Assumes constant comparison between countries and time with respect to age and structure without the effect of calculating age distribution with a population

As per Institute of Health Metric and Evaluation the age group between 15-24 tends to have more suicidal tendency than any other age group and it has a gradual drop from late 90's to 2000's.In 2016 there was an estimate of 817800 suicides and there was a small reduction from 90's where annual suicide count was around 850000-860000

2.4 Gender Difference in Suicidal Rates

It is a fact that , when we consider cross-country figures, it does not show any clear trends and the characteristic of suicidal behaviour changes and varies widely between different countries, communities in different economic background over time. If gender and scio economic roles contribute to sex variations inside the suicide rate, the count would be on the significance of the sex distinction over time in accordance with temporal modifications in gender equality in function career. In the beginning look, the facts do not seem to help this. said that the movement of girls into the paid labour pressure within the 1970s did not increase the woman suicide price, even amongst those girls who juggled their employment with being a wife and mother. Conversely, there was a better male suicide rate at some stage in the Nineteen Seventies in communities with a better price of female labour pressure participation amongst married girls with small children. One interpretation is that guys have been unable to deal with the loss of their role as the only issuer, resulting in a loss of self esteem amongst guys, specifically individuals who adhered most strongly to standard gender roles . Any contribution of gendered social-function career to suicide price, then, is not likely to be a simple additive effect, and could have interaction with prevailing norms, the volume to which a society has embraced gender equality, and local social ecology. Certainly, this is tested by way of special outcomes of female hard work pressure participation at the suicide rates of males and females in Canada between 1971 and 1981 .In 1971, whilst prevailing societal views of married women in paid employment were in large part terrible, girl exertions force participation elevated suicide danger for ladies and men. by using 1981, but, with greater recognition of women in paintings, girl labour pressure participation reduced the risk of suicide for both sexes.

The ratio of suicide from male to female varies from country and region. As per a research ,it states that , by comparing the female suicide rat, male weigh five times higher in Europe , 3.6 times in the United States , 3 times in the United Kingdom and 2.5 times in the Korean Republic, and a study conducted in Turkey found that average rate of suicide in the year 1987 and 2011 for male was 1.8 times higher than that of the females but as fact male to female ratio is lower in turkey

As we go through the recent years between 2014 and 2015, in the United Kingdom , male suicide decreased and female suicide increased since 2014, but when we peek a look into Ireland and Scotland , both male and female suicide decreased on time, and northern Ireland , Wales and in the United States there were a booming trend of people from both sex committing suicide . It is noticed that there is gender difference in suicide tendency in different nation and in different communities.

Suicide rate depends on the head count and due to difference in head count , only looking at the rate of suicide may be misleading , so to control the population effect difference in specific groups, the standardized method was used to specify gender wise suicides and it was done by dividing number of suicides by each gender by he corresponding head cont in that particular gender and been multiplied by hundred thousand . The above mentioned calculations was used to calculate gender wise suicide separately for different years.

2.5 Suicide in Dundee– Gender Based Analysis

The hypothesis is been tested whether the movement of working class women that too in the paid labour affects the gender difference in the suicide rate , in places where working class women bring bread to home. The tested data has been extracted from the newspaper report of completed and attempted suicide rate for the Scotland city, Dundee , in the nineteenth century and early twentieth century. The particular place and time has been chosen because ,it provides a unique opportunity to explore the evolution from an almost male dominant working labour to female bread winners in under fifty years. The City's jute mill industry consider women as their primary work force , as their wages were less compared to the male and there were lesser number of women working and eventually Dundee began to know as ‘She’ town. Exploratory data analysis were done to determine whether gender plays an important role, from the time where women become a huge work force in the city. To know the sex difference between male and female , in attempted an fully completed suicides ,for that they compared and calculated number of suicides per hundred thousand individual of the female and male population . Male and female head counts or the population were obtained from census and after statistically analyzing the data , it was found that 144 attempts were done by women abd 127 women dies by suicide and in the case of male 178 attempts and 250 killed themselves.

2.6 Economic Recession

Various studies have analyzed the link between economic recession and suicide rates. 9 Death Statistics were used in the WHO and Taiwan to examine whether there was a link between the Asian economic crisis of 1997-1998 and the rate of suicide. The following charts summarize their results. As we can see, the rates of men in 1998 increased significantly in Japan, Hong Kong and Korea, while the increase in women's rates was less pronounced in the same countries. There were no similar patterns of suicide rates in Taiwan and Singapore, where the economic crisis had less impact on the economy. Researchers such as McKee , Reves and Skutler in 2014 found these similar results and they found there were 21 people die per hundred thousand individuals. Most of the research studies were focusing on trends and patterns of effect of economy on a particular place regarding a specific event at specific nations or regions .The studies were done using cross country data once different nations ,on a long span of time and they got similar results.

Lately , the analyze was focused on public data on suicide , population and economic background on a period of 2000-2011 into 63 nations as in 4 different regions and the results shows that rise in unemployment resulted in more suicidal tendency amo9ng individuals living in those particular places, but they found that relationship is non linearly co-related between unemployment and suicides

2.7 Research done in china

A research was done in Republic of China to see whether unemployment is causing suicidal tendency on p[people .This data was comprised of economic index during 2004-2013 and were collected from National Bureau of Statistics of China. This data includes Gross Domestic Production per capitol and the regional level data included , GDP per capitol in 3 different regions (west, central and east) , six areas annual income per capitol(East urban ,central urban, west urban, east rural, central rural, west rural) all were manipulated for expansion. The researchers created a new index and called VCMI, which can calculate annual fluctuation measure of the stock market

VCMI= Daily closed price mean/daily closed price’s standard deviation

This economic index was presented by RMB(Chinese currency) and at hat time US$=6.35 RMB

Gross domestic production is the sum of gross value summed by all nation’s productions on the market and added to their product taxes subtracted their subsidies . It is determined without making any deduction for assembled assets. Gross domestic production per capitol was defined as the GDP on mid year population. The yearly salary per capitol alludes to part of the income of all family money pay that can be utilized to organize the family life. Spellbinding insights were displayed, including yearly ratter of suicides and suicide rate per hundred thousand people for every one of the multiyear ag, gender, and area explicit gatherings. Across the country and district level GDP and annual salary per capitol were altogether estimated from the Chinese cash RMB and balanced by swelling. Nonparametric Kruskal– Wallis ANOVA was utilized to test arithmetical contrasts among three districts. As the information were too scattered to even think about using Poisson relapse, the creators picked negative binomial relapse to analyze the adjustment in the district level suicide rates after some time and the impact of age, sex, locale, and monetary record on the suicide rate. The impact of age, sex, district, and financial list on the suicide rate was dissected utilizing negative binomial relapse model,26 with the stratified suicide number as the reaction variable, the characteristic logarithm of stratified DSPs' populace as a counterbalance, and the multiyear age-gatherings, sex, urban/country territory, and monetary file as the logical factors. Time patterns of suicide rate in various zones were likewise dissected utilizing negative binomial relapse, with the stratified suicide number as the reaction variable, the normal logarithm of stratified DSPs' populace as a counterbalance, and year as an illustrative variable. To diminish the variety, GDP per capita in this investigation was changed over to a unit of 1,000 RMB, and after that changed by the characteristic logarithm. These analysis were generated using SPSS version 21. All were reporting probabilities (value as p) were 2 sided and those were less than 0.05 were significantly statistical.

Results of this research was notable, it was found that country wise and region wise gross domestic production per capitol were tremendously increased annually , whereas VCMI posted a peak in 2008, GDP per capitol and yearly urban and rural salary per capitol had almost the same trend , it showed a high peak in East area and least in West side , VCMI boosted largely around 2008 and decreased after that time.

2.8 Insight on Korean economic background and Suicide rates

The greater part of amasses in the psychological composing have advanced toward the examination of danger factors and partners of foolish practices from a clinical perspective in made countries. In these countries, clinical examinations definitively set up that psychological dissipates are an imperative contributing segment to suicide. It isn't absolutely clear how well these examinations can be extrapolated to the comprehensive network that psychological disarranges are tolerably uncommon in the general open and a large portion of subjects dissected as having a mental issue don't complete suicide. The impact of mental peril factors on suicide rates changes in a sense that depends upon money related status in Western made countries and, even more basically, this impact may be significantly humbler in making countries.

Researchers Morseli and Duchem are the first researchers to put up the socioeconomic factors which pulls on the changing suicide rates. ince their novel nineteenth century contemplates, suicide has as a rule been viewed as a social problem, and several danger factors have been identified with inconsequential immediate, both at the subject estimation (microsocioeconomic factors) and at the state level (macro socioeconomic level). Beginning late, Inamoratas et al8 low down, constantly with past research, crucial association between suicide rates and a couple macro socioeconomic factors. Both high basal components of remuneration and raised financial improvement have been believed to be associated with diminished suicide rates. Suicide rates are lower in Western high-remuneration nations separated and low-and centre pay nations. Also, the general change in cash related exercises that has occurred over the past 60 years has altogether influenced individuals' energized success and may have influenced suicide rates.11 Gross private thing (GDP), a degree of monetary action, has been oppositely connected with suicide rates, proposing that suicide rates drop in the midst of budgetary headway and enlargement in the midst of recession.12 beginning late, prescribed that business cycles sway suicide rates in the USA: the general suicide rate when in doubt moves amidst retreats and falls amidst developments. Thusly, the Asian money related emergency has been related with retain builds suicide rates in a couple, yet not all, East/Southeast Asian nations.

It has been endorsed that the relationship between GDP per capita and suicide may look for after an exchanged U-melded bend, with suicide plans declining subsequent to cresting at a specific edge of financial movement. In like way, paying little heed to the path that at low GDP levels, enlarges in GDP are associated with increases in suicide rates, when a given edge of cash related movement is cultivated, further expansions in GDP don't interface with further growths in suicide rates. The edge at which the modified U-framed bend begins inclining down may change subordinate upon two or three social, money related and social separations crosswise over nations. Obviously, two or three producers have recommended that, separated and GDP, progressively complex structures, for example, 'National Intelligence' or the 'K factor'— including degrees of national IQ; net national thing; future; birth-rates; youngster mortality; HIV/AIDS and of assault, authentic strike and murder—may better clear up the geographic impulse of suicide rates.

In the present examination, we evaluated the models in GDP per capita balanced for acquiring power consistency (PPP) and suicide rates and investigated the association between's PPP-balanced GDP per capita and suicide rates in 10 WHO areas amidst the past 30 years. The present examination develops past research by including each and every accessible datum around the world, all things considered permitting refinement and light of the provincial complexities that may clear up the relationship between cash related cycles, as surveyed by PPP-balanced GDP per capita, and suicide rates.

2.9 Population and Suicide

There is a process called apoptosis where cells commit suicide, it occurs when a cell from clusters lost it’s pack and self destruct itself and a biologist Matt Roft said that its the other cells which makes them to live in a pack and the isolated one die for the others to survive .This study can be compared to human beings as well . When in a group it is stronger and when one it gets isolated it destroys themselves (KHAZAN, 2014)s. A research paper states that as the density of population increases there is a lesser chance of suicide to happen .

But a researcher from china name Wang found that death increased as the density of population increased when they filtered age group 15 -29.

As more people moved into the cities about one more factors changes and increases and one among this is the death toll (JAFFE, JUL 15, 2013). When they compared Brazil and the United States during the year 1992 to 2009 they found that rate of suicide boomed slower than population by 76% ,when the head count of people went up to 100. As per them , they came into conclusion that the suicide depends on an individual choice , most people who commits suicide wants to live but as they lives in he big cities , feel lonely a they won't be having any shoulder to depend up on , to share their feelings and share their problems , and they don't have any way to escape other than killing themselves and they concluded that population doesn't affects or depends the suicide rates.

3. Data Analysis

3.2 Machine Learning

Machine learning is a way of statistics analysis that automates analytical version building. It is a department of artificial intelligence based totally on the idea that structures can research from statistics, identify patterns and make choices with minimum human intervention. Because of the new generation technologies , machine learning nowadays it has changed a lot from its past techniques and algorithms. It has given birth to system which can learn patterns and algorithms and do tasks and learn like a human being without being programmed to perform a certain task. Researchers were interested whether AI (Artificial Intelligence) enabled computers can learn from data. The complete aspect of machine learning is considered because , as algorithms or methods were opened to new data , they are able to adapt its value and compute accordingly. They act like human brains to learn from previous activities and generate new data. The growth in advance machine learning algorithms have influenced and simplified data analytics and forecasting in big data. it is seen as a subgroup of AI (Artificial Intelligence) .They are widely used in variety of places to perform different kinds of tasks .They mainly focuses on making predictions using system and the data it holds. There are mainly two types of machine learning algorithms , supervised and unsupervised machine learning algorithms

Supervised machine learning is the greater commonly used between the two. It consists of such algorithms as linear and logistic regression, multi-class classification, and guide vector machines. Supervised studying is so named due to the fact the facts scientist acts as a guide to train the algorithm what conclusions it must come up with. It’s similar to the way a child might analyse arithmetic from a teacher. Supervised learning requires that the algorithm’s viable outputs are already known and that the facts used to educate the algorithm is already labelled with right answers. For example, a classification algorithm will learn to become aware of animals after being trained on a dataset of photos that are true labelled with the species of the animal and some identifying characteristics.

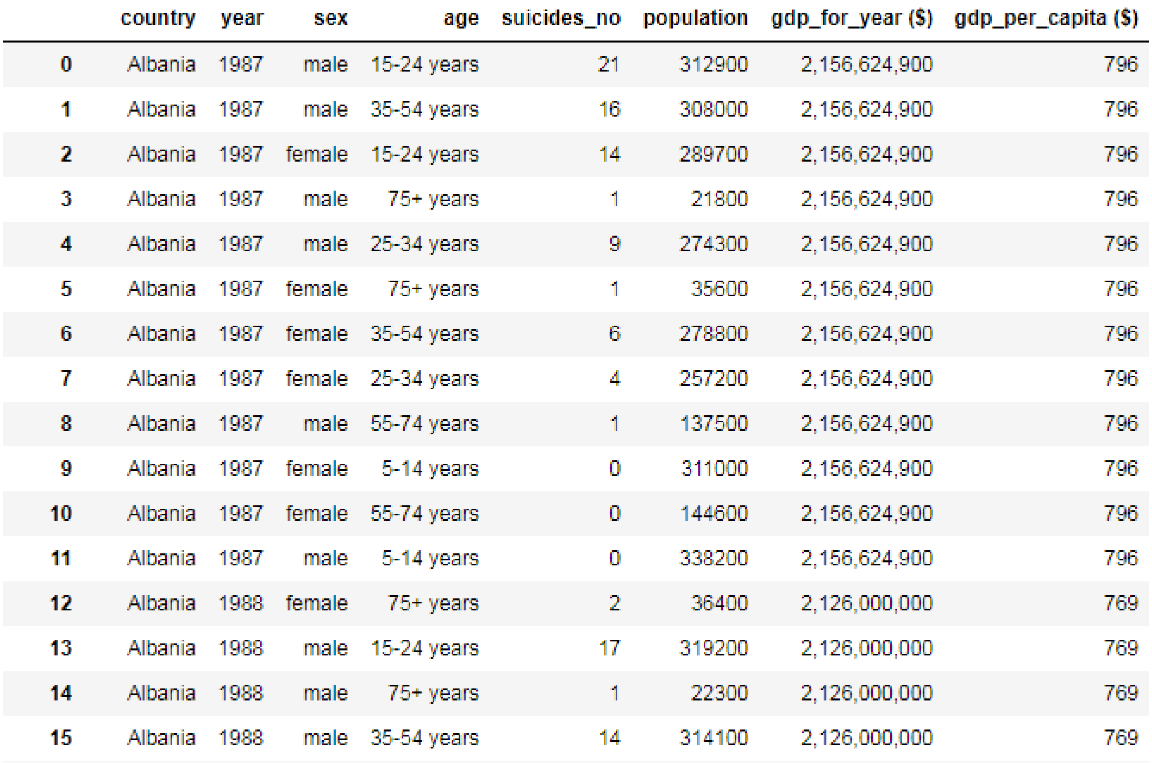
On the other hand, unsupervised machine learning is extra closely aligned with what some call true synthetic intelligence — the thinking that a computer can examine to pick out complicated processes and patterns except a human to grant training alongside the way. Although unsupervised gaining knowledge of is prohibitively complicated for some easier organisation use cases, it opens the doorways to solving troubles that people typically would no longer tackle. Some examples of unsupervised laptop learning algorithms include k-means clustering, fundamental and impartial issue analysis, and affiliation rules.

While a supervised classification algorithm learns to ascribe inputted labels to pics of animals, its unsupervised counterpart will appear at inherent similarities between the photographs and separate them into organizations accordingly, assigning its personal new label to each group. In a realistic example, this type of algorithm is useful for purchaser segmentation due to the fact it will return corporations based on parameters that a human can also now not reflect on consideration on due to pre-existing biases about the company's demographic.

Choosing to use both a supervised or unsupervised machine learning algorithm commonly relies upon on elements associated to the structure and extent of your statistics and the use case of the difficulty at hand. Well rounded information science software will use each sorts of algorithms to build predictive information models that assist stakeholders
make decisions throughout a variety of enterprise challenges.

3.3 Dataset

The data set is an overview of world suicide rate from 1985 to 2015 , which compares socio-economic features with suicide rats with year and country. This accumulated dataset pulled from four distinctive datasets associated by time and place, and was attempted to find signals related to extended suicide rates among different accessories all around, over the money related range. The dataset comprises of attributes from country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, HDI for year, gdp_for_year, gdp_per_capita, generation (based on age grouping average). HDI for year column has been dropped as it had null values as 75% of its column, we removed suicide/100k population , that is exactly the value which we get when we divide population by suicide, then we removed 'country-year' , that is the same column as 'country' and 'year', then 'generation type' is removed as we csn compare them with the age group and the filtered dataset looks like this:

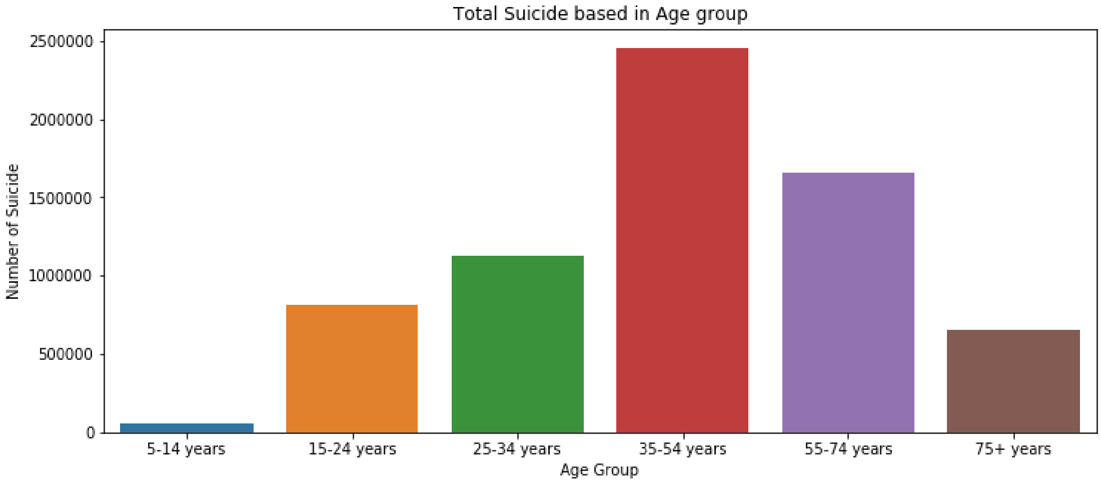


	country	year	sex	age	suicides_no	population	gdp_for_year ($)	gdp_per_capita ($)
0	Albania	1987	male	15-24 years	21	312900	2,156,624,900	796
1	Albania	1987	male	35-54 years	16	308000	2,156,624,900	796
2	Albania	1987	female	15-24 years	14	289700	2,156,624,900	796
3	Albania	1987	male	75+ years	1	21800	2,156,624,900	796
4	Albania	1987	male	25-34 years	9	274300	2,156,624,900	796
5	Albania	1987	female	75+ years	1	35600	2,156,624,900	796
6	Albania	1987	female	35-54 years	6	278800	2,156,624,900	796
7	Albania	1987	female	25-34 years	4	257200	2,156,624,900	796
8	Albania	1987	male	55-74 years	1	137500	2,156,624,900	796
9	Albania	1987	female	5-14 years	0	311000	2,156,624,900	796
10	Albania	1987	female	55-74 years	0	144600	2,156,624,900	796
11	Albania	1987	male	5-14 years	0	338200	2,156,624,900	796
12	Albania	1988	female	75+ years	2	36400	2,126,000,000	769
13	Albania	1988	male	15-24 years	17	319200	2,126,000,000	769
14	Albania	1988	male	75+ years	1	22300	2,126,000,000	769
15	Albania	1988	male	35-54 years	14	314100	2,126,000,000	769

3.4 Exploratory Data Analysis

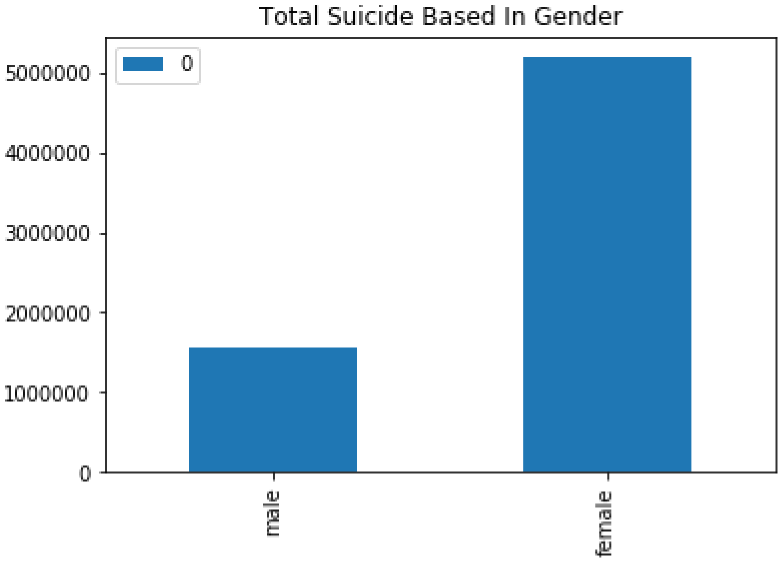
The initial investigation of data starts when one understands data completely and it is easier to understand a data when one sees it it visual , here are some facts regarding the suicide data

3.4.1 Age Group Visualization



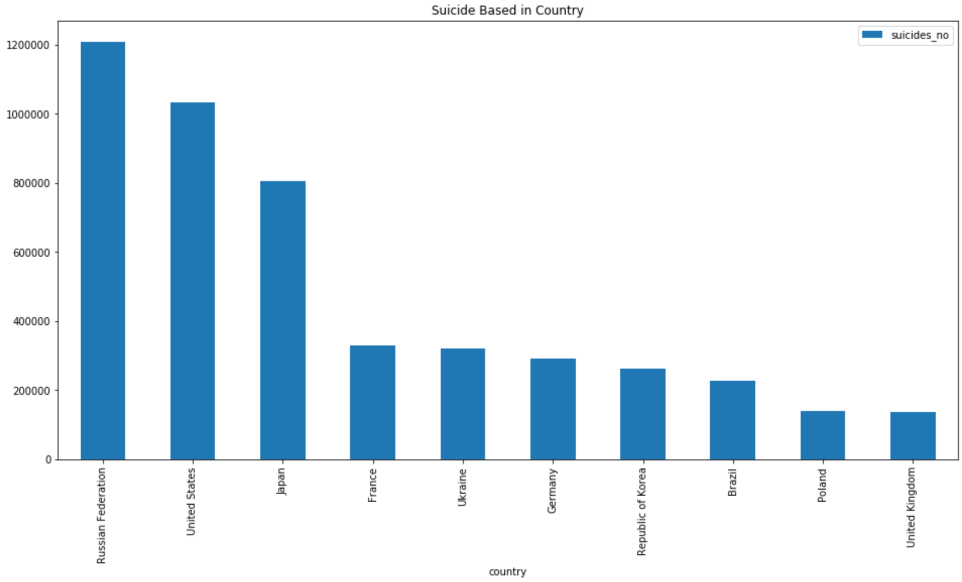
According to the data it is clear from the graph that age group from 35-54 has got the maximum suicidal tendency and during the years they accounted around 23 hundred thousand suicides

3.4.2 Suicide gender wise



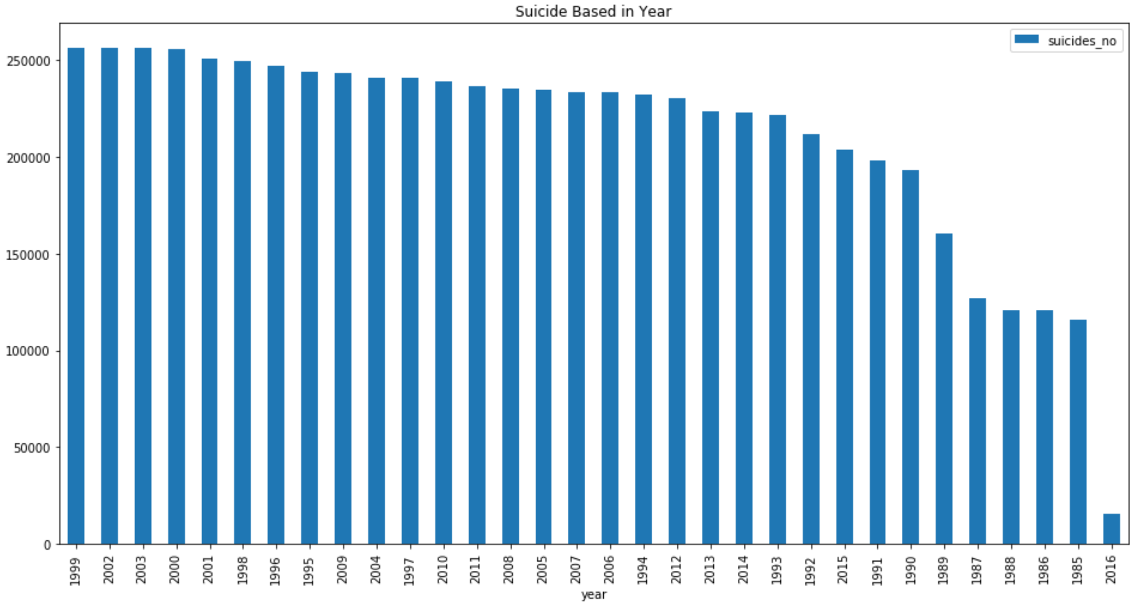
As far the sex is considered the females have committed more suicides in these years and accounted more than fifty hundred thousand

3.4.3 Suicide –Country Wise



According to the graph above the Russian has accounted for the maximum suicide followed by the United states and then Japan.

3.4.4 Suicide Rate - Year Wise



According to the data it shows during the years 1999,2002,2003 and 2000 , the suicide rate peaked the maximum and it had a gradual decline as we go down through the years.

3.5 Linear Regression

Linear regression is a supervised machine learning algorithm which perform task for regression. This model for regression predicts the independent variable by considering the target variable. Linear regression is commonly applied to find out the relation between variable and forecasting . Dependent and independent variable differ based on their different regression, number of independent variables are being used are considered. The task of linear regression is to predict a dependent variable named (Y) on the basis of an independent variable named (X) and their regression calculate the relationship between X and Y and thus called as linear regression (berlanga, n.d.) .

Function Hypothesis

$Y=\theta 1+ \theta 2.X$

X is the training data

Y is the label to the data

θ 1 is the intercept

$\theta 2$ is the co-efficient of X

3.6 Decission Tree

Decision tree is like a flowchart tree structure , in which each nodes internally donates an attribute test , outcoem of the test is represented in branches anad moreover class labels are hold by leaf nodes.They are able to produce meaningful results and evaluates classification without much computation . They are made to handle both continous and categorical variables . They give a correct indicaton of important prediction or classification fileds.Its primary task is to support analysis to reach target variable (tree, n.d.).

3.7 Support Vector Machines

Support Vector Machines is one among the supervised algorithm which is commonly used for regresssion or classifiaction .Based on transformation , it finds an excellent branching for possible outputs (quantstart, n.d.). Support Vector Machines generates the model by creating a feature space, the main aim of SVM is to create a new empty object by training a model, this is being achieved by creating a partition of feature space. It is mostlyy used in sentimental analysis and classification ,in short they are high dimensional , they are memory efficient , as it only have one subset of the train set are used. They are versatile in class seperation (kdnuggets, n.d.).

3.8 Random Forest

Random forest is a supervised algorithm which creates multiple decision tree and blend them together to gt a more decent an accurate predictions. There is a technique in randowm forest called bagging , it involves training different data in decision tree where sampling and replacement are done together. Random forest tends to work accurately with a large number of datas. They arehaving a high accuracy ratio and are flexible . Random forest doesn’t over fit the data.

3.9 K-Nearest Neighbour

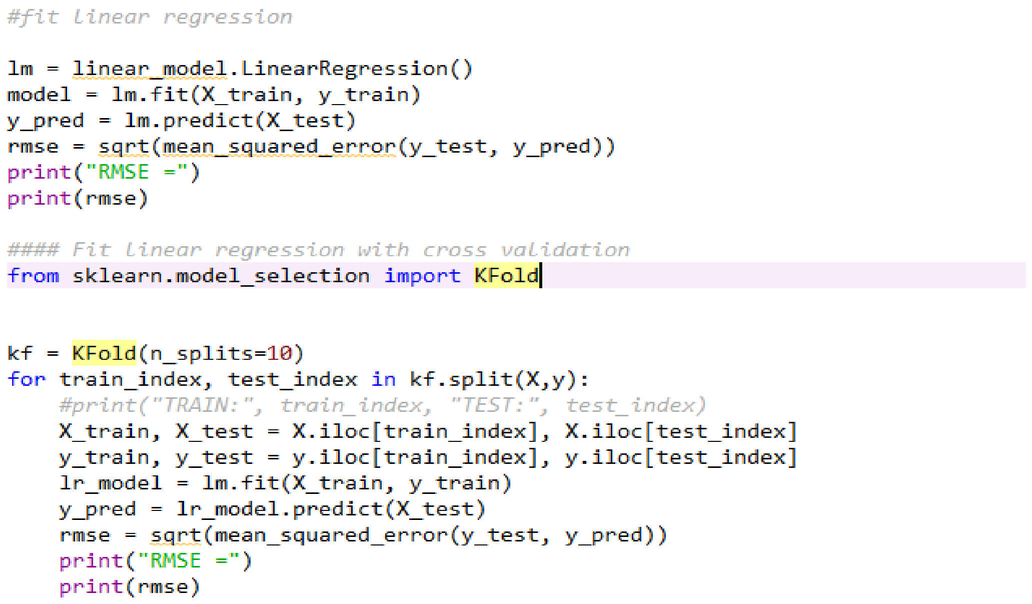
K Nearest Neighbour algorithm is a type of instance based learning algorithm (algorithm, n.d.). K-Nearest Neighbour is one among the most common classification algorithm used to classify and represen data. They use similar measures to use a data and measure its data points, majority votes by its neighbour is used to classify data. The most nearset neighbouring is asigned to the data. K value accuracy increases as the number of nearest neighbour increases , it is used when the outputs are easy to intercept . In KNN the property valuefor the object is the output (Anon., n.d.).

4. Results

As we tying to predict the suicide number by using suicide number as the target variable by applying five of the machine learning algorithms to compare their RMSE value and predicting which algorithms works better for the dataset. There were limitations as we doesn’t had a big dataset ,as the dataset only contained twenty eight thousand rows ,it was very difficult to find an accurate score.

The following are the RMSE scores for each algorithm with their model fitting

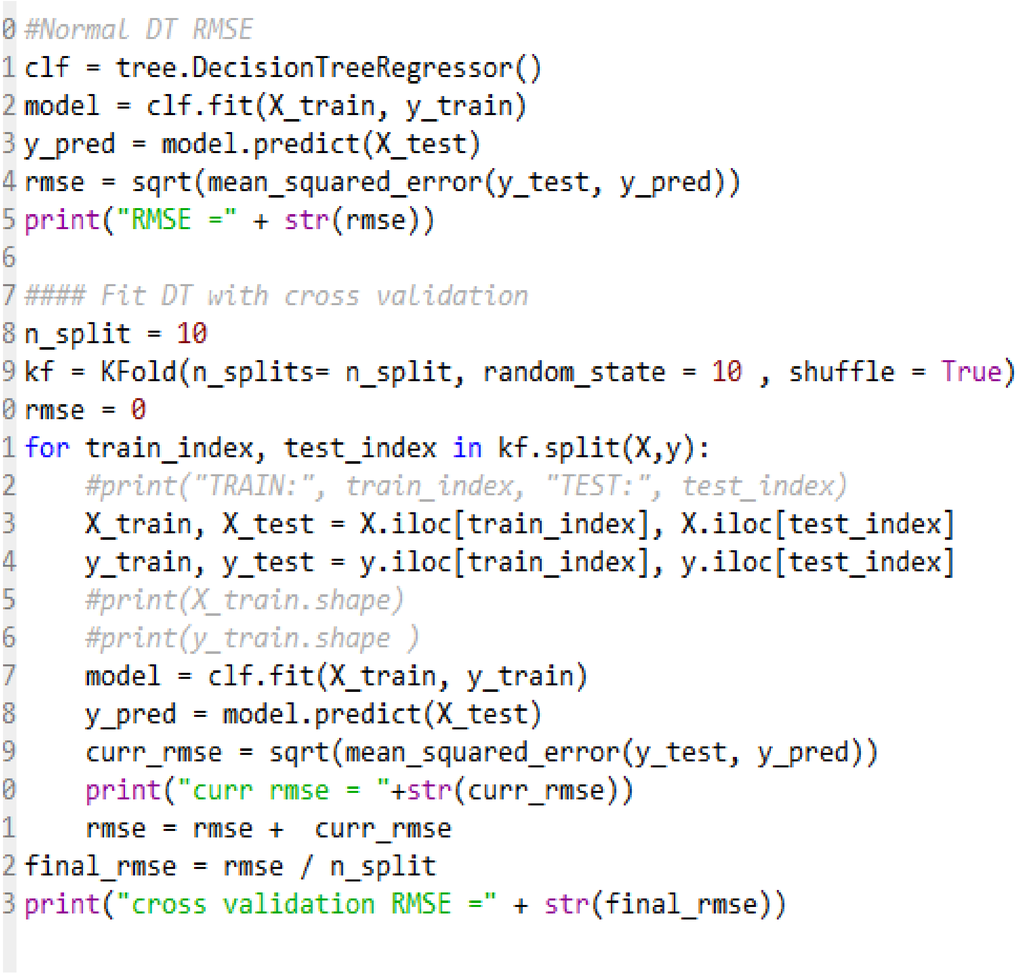
4.1Linear Regression



```
#fit linear regression

lm = linear_model.LinearRegression()
model = lm.fit(X_train, y_train)
y_pred = lm.predict(X_test)
rmse = sqrt(mean_squared_error(y_test, y_pred))
print("RMSE =")
print(rmse)

#### Fit linear regression with cross validation
from sklearn.model_selection import KFold


kf = KFold(n_splits=10)
for train_index, test_index in kf.split(X,y):
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
    lr_model = lm.fit(X_train, y_train)
    y_pred = lr_model.predict(X_test)
    rmse = sqrt(mean_squared_error(y_test, y_pred))
    print("RMSE =")
    print(rmse)
```

The cross validation RMSE score we got after implementing linear regression was **639.1736520412852**

4.2.Decision Tree

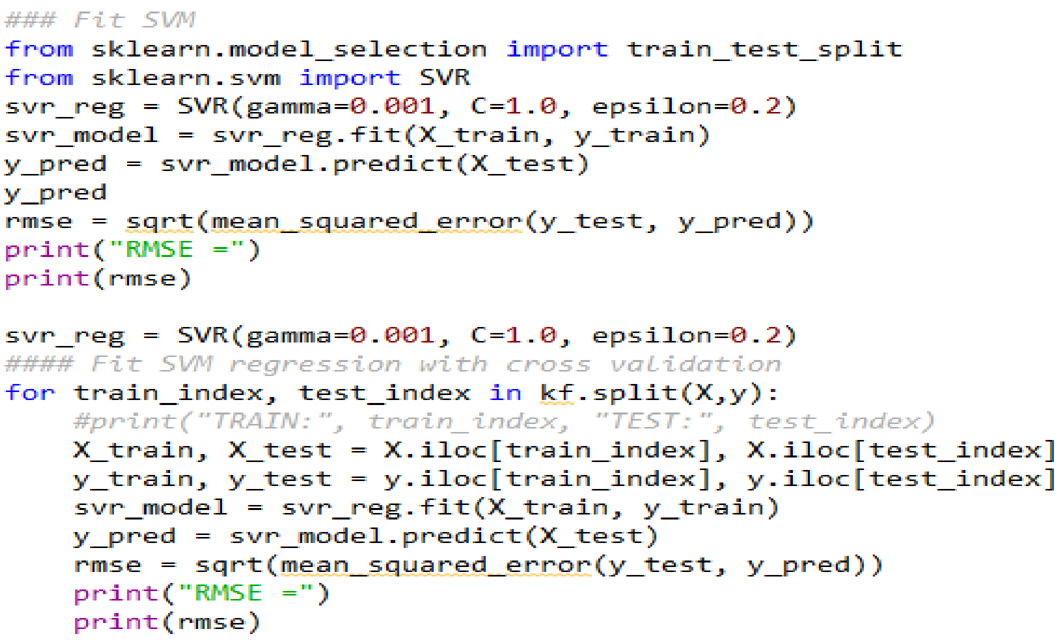


```
#Normal DT RMSE
clf = tree.DecisionTreeRegressor()
model = clf.fit(X_train, y_train)
y_pred = model.predict(X_test)
rmse = sqrt(mean_squared_error(y_test, y_pred))
print("RMSE =" + str(rmse))

#### Fit DT with cross validation
n_split = 10
kf = KFold(n_splits= n_split, random_state = 10 , shuffle = True)
rmse = 0
for train_index, test_index in kf.split(X,y):
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
    #print(X_train.shape)
    #print(y_train.shape )
    model = clf.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    curr_rmse = sqrt(mean_squared_error(y_test, y_pred))
    print("curr rmse = "+str(curr_rmse))
    rmse = rmse +  curr_rmse
final_rmse = rmse / n_split
print("cross validation RMSE =" + str(final_rmse))
```

The cross validation RMSE score we got after implementing Decision Tree Regressor was

186.40665183692244

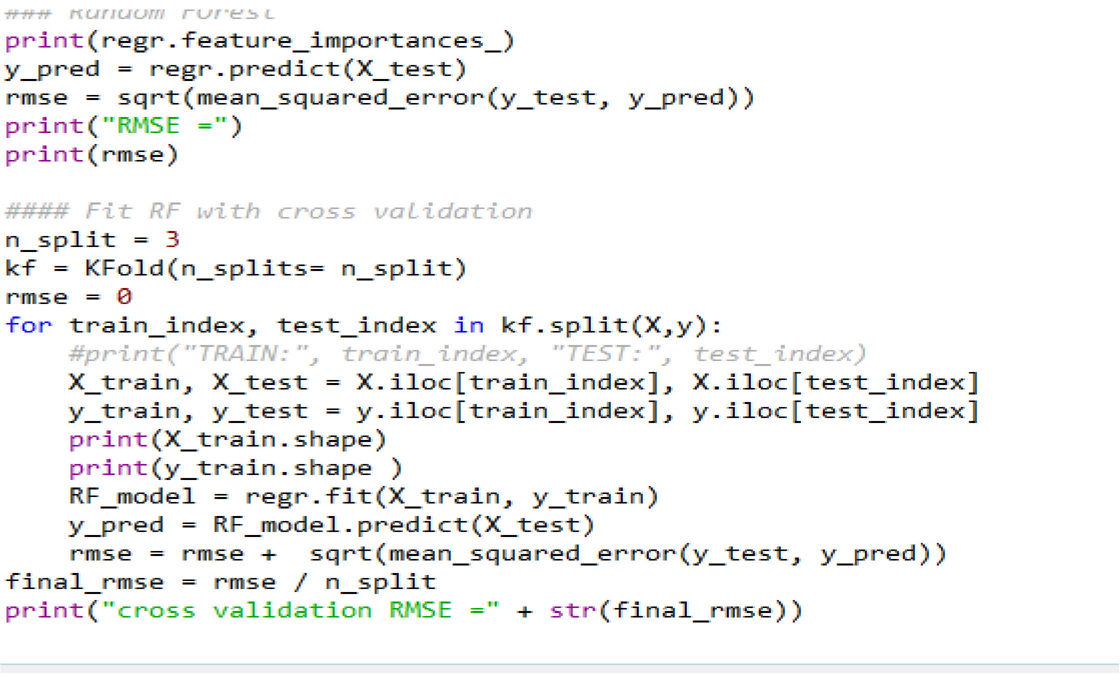
4.3Support Vector Machines



```
### Fit SVM
from sklearn.model_selection import train_test_split
from sklearn.svm import SVR
svr_reg = SVR(gamma=0.001, C=1.0, epsilon=0.2)
svr_model = svr_reg.fit(X_train, y_train)
y_pred = svr_model.predict(X_test)
y_pred
rmse = sqrt(mean_squared_error(y_test, y_pred))
print("RMSE =")
print(rmse)

svr_reg = SVR(gamma=0.001, C=1.0, epsilon=0.2)
#### Fit SVM regression with cross validation
for train_index, test_index in kf.split(X,y):
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
    svr_model = svr_reg.fit(X_train, y_train)
    y_pred = svr_model.predict(X_test)
    rmse = sqrt(mean_squared_error(y_test, y_pred))
    print("RMSE =")
    print(rmse)
```

The cross validation RMSE score we got after implementing Support Vector Machine was **595.7227269763561**

4.4Random Forest

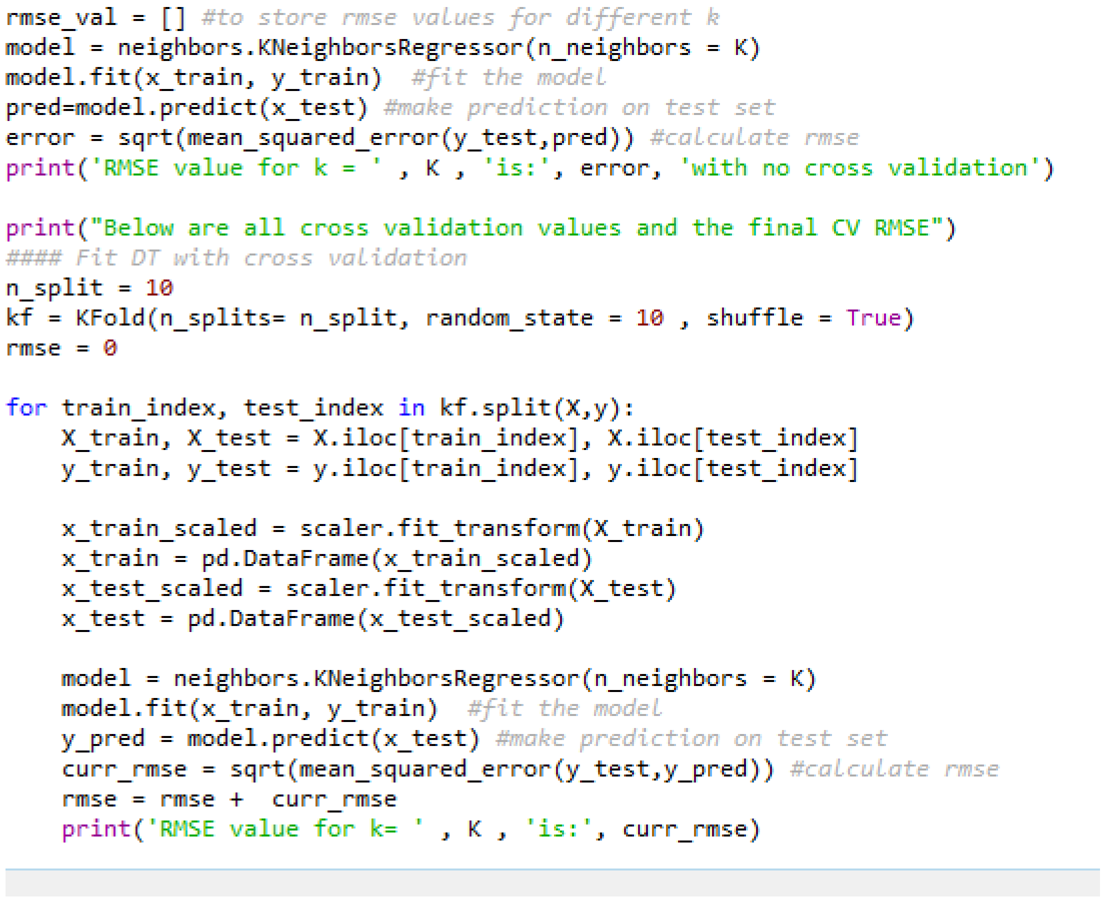


```
### Random Forest
print(regr.feature_importances_)
y_pred = regr.predict(X_test)
rmse = sqrt(mean_squared_error(y_test, y_pred))
print("RMSE =")
print(rmse)

#### Fit RF with cross validation
n_split = 3
kf = KFold(n_splits= n_split)
rmse = 0
for train_index, test_index in kf.split(X,y):
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
    print(X_train.shape)
    print(y_train.shape )
    RF_model = regr.fit(X_train, y_train)
    y_pred = RF_model.predict(X_test)
    rmse = rmse +  sqrt(mean_squared_error(y_test, y_pred))
final_rmse = rmse / n_split
print("cross validation RMSE =" + str(final_rmse))
```

The cross validation RMSE score we got after implementing Random Forest was

793.5550378944887

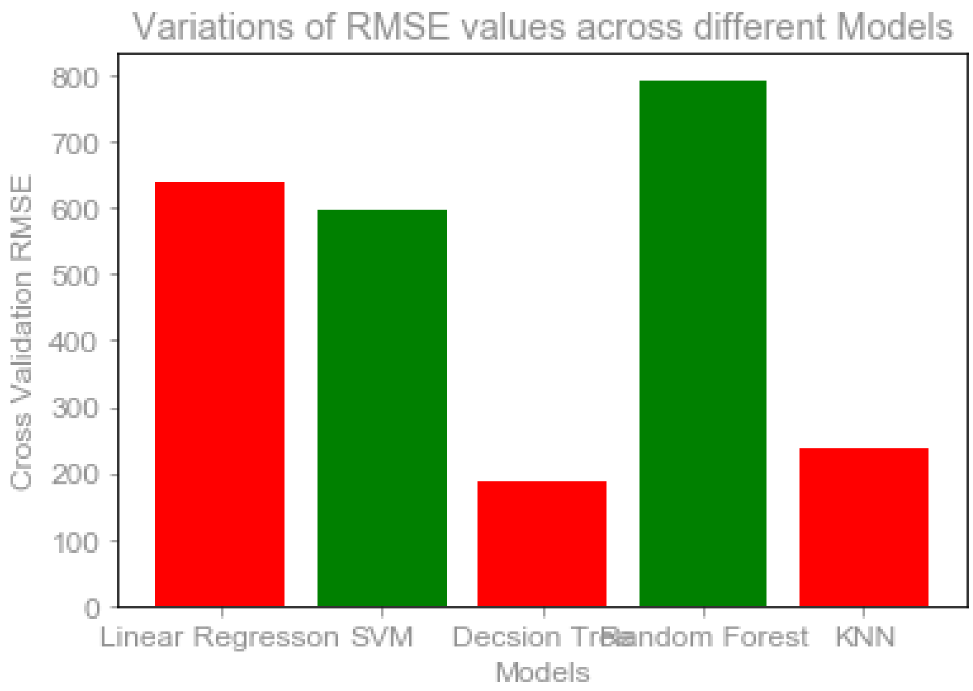
4.5K- Nearest Neighbour



```
rmse_val = [] #to store rmse values for different k
model = neighbors.KNeighborsRegressor(n_neighbors = K)
model.fit(x_train, y_train)  #fit the model
pred=model.predict(x_test) #make prediction on test set
error = sqrt(mean_squared_error(y_test,pred)) #calculate rmse
print('RMSE value for k = ' , K , 'is:', error, 'with no cross validation')

print("Below are all cross validation values and the final CV RMSE")
#### Fit DT with cross validation
n_split = 10
kf = KFold(n_splits= n_split, random_state = 10 , shuffle = True)
rmse = 0

for train_index, test_index in kf.split(X,y):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]

    x_train_scaled = scaler.fit_transform(X_train)
    x_train = pd.DataFrame(x_train_scaled)
    x_test_scaled = scaler.fit_transform(X_test)
    x_test = pd.DataFrame(x_test_scaled)

    model = neighbors.KNeighborsRegressor(n_neighbors = K)
    model.fit(x_train, y_train)  #fit the model
    y_pred = model.predict(x_test) #make prediction on test set
    curr_rmse = sqrt(mean_squared_error(y_test,y_pred)) #calculate rmse
    rmse = rmse +  curr_rmse
    print('RMSE value for k= ' , K , 'is:', curr_rmse)
```

The cross validation RMSE score we got after implementing K-Nearest Neighbour was

237.57899678664558



So, the decision tree is the best and suitable algorithm to predict the suicide number from this dataset.

5. Conclusion

In this paper, well-known machine algorithms, Linear Regression , Decision tree , Support Vector Machines, Random Forest and K- Nearest Neighbour is being applied on the dataset for predicting the suicide number and identifying which algorithm works better for the data set. The aim of this thesis is to predict the suicide number in each country by considering certain factors influencing the people who are affected by it and helping the community and government to forecast crisis and problems ahead in the future and take measures and steps to tackle them wisely . The suicide dataset around the world was used , that gives a data from 1985-2015 and five algorithms were applied separately . Each of the algorithms has its own properties and the best score was given by the decision tree.

4.6The limitation for predicting suicide

The pleasing potential outcomes for world suicide bar do exclude obsolete contemplations of suicide figure or peril examination. diminishing the suicide cost in countries with a high suicide run after or conveying the suicide charge of partners to nearer that of young women would possibly get monstrous declines in by and large suicide rates. thought in regards to the thought of hospitalized sufferers can not be unnoticed in light of the striking suicide cost on this affiliation, anyway this may least troublesome be depended upon to have an unassuming result on all things considered suicide costs because of the very reality most suicides are with the help of individuals United Nations association have not been in an amazingly remedial distinguishing strength sanatroium. additional for the most part, restorative distinguishing strength treatment should be given to absolutely everybody thusly on decrease their weight of signs and signs and wish to now not be appropriated or even by suggests that of contemplations of United Nations office is likely or not visiting suicide. at the unclear time as specific patients can create more prominent issue basically suicide than others, perception of the obliged affectability, unassuming force of division, and besides the horrendously low gainful prophetic cost of suicide risk evaluation should energize clinicians inside the undertaking of joint assurance making with their absurd sufferers.

Disregarding the impulse to envision suicide in clinical helpful specialty seek after may even empower suicide bar. Low helpful prophetic characteristics infer that the general open United Nations association get fix because of a predominant hazard course of action can by no means, that fail horrendously by maltreatment suicide and besides the restricted affectability approach that as very nearly 1/2 of the sufferers who do kick the can by suicide may require been underprivileged of assurance measures once a lower-chance request. epistemological choices practically future suicide ought to be made horrifyingly completely and best paying little heed to the to be had affirmation is extended. considerable associated math hazard portions may produce a promise to such Associate in Nursing epistemological name regarding sociality, in any case this dedication should be unpretentious.

Instead of endeavour and build a suicide desire, clinicians should mindfulness on redesigning the collaboration with the patient in case you wish to develop trust, reduce the patient's wretchedness and suffering, and expand the patching agreement. A total assessment of the patient's propelled needs should adapt to. these needs can overtimes passes on with it the essential to oversee modifiable portions that are related to suicide, as Associate in Nursing point of reference treatment of substance use, anyway most such needs should be met paying little personality to the association with destiny suicide. needs examinations aren't probabilistic and should cause fixes being offered to any or all patients paying little mind to saw suicide peril.

4.7Future Scope

In this thesis paper, we were trying to achieve the best possible machine learning algorithm for predicting the suicide number by taking some influential factors and variables . This field in suicide prediction is like a vast ocean where big possible predictions can be made to stop the out growing trend of suicide among societies and around the world .A very advanced deep learning model can help the people and the government to forecast a crisis and disaster that might be an economic situation where people are not ready to tackle, this might lead them to a mental breakdown and generates suicidal tendency among them , so one such can help everyone to avoid a hara-kiri situation and make a world a better place to live.

Reference

A, B., 2013. *Violence risk assessment in clinical settings: being sure about being sure,* s.l.: google scholar.

algorithm, k.-n. n., n.d. *wikepedia.* [Online]
Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Anon., 1997. In: *Suicide as a cultural institution.* russia: cornell university press.

Anon., 2017. *Annual Research Review: Suicide among youth – epidemiology, (potential) etiology, and treatment,* s.l.: https://doi.org/10.1111/jcpp.12831.

Anon., n.d. A SURVEY PAPER ON SUICIDE ANALYSIS. *International Journal of Pure and Applied Mathematics,* Volume 8.

Anon., n.d. *pioneer institute in algo training.* [Online]
Available at: https://www.quantinsti.com/blog/machine-learning-k-nearest-neighbors-knn-algorithm-python

Anon., n.d. *Suicidal speech essay,* s.l.: study moose.

berlanga, f., n.d. *geeks for geeks.* [Online]
Available at: https://www.geeksforgeeks.org/ml-linear-regression/

Burnap P, C. W. S. J., 2015. *Machine classification and analysis of suicide-related communication on twitter.,* s.l.: Proceedings of the 26th ACM Conference on Hypertext & Social Media.

Caruso, K., n.d. Suicide is NOT Self-Murder.

Ceccherini-Nelli A, P. S., 2011. *Economic factors and suicide rates: associations over time in four countries.,* s.l.: Soc Psychiatry Psychiatr Epidemiol.

Center, S. P. R., n.d. *SPRC.* [Online]
Available at: https://www.sprc.org/about-suicide/scope

Gregory, C., n.d. *Suicide and Suicide Prevention,* s.l.: psycom.

Health., N. I. o. M., 2008. *Suicide and Suicidal Behavior,* s.l.: National Institute of Mental Health..

JAFFE, E., JUL 15, 2013. *The Unsettling Link Between Sprawl and Suicide,* s.l.: city lab.

kdnuggets, n.d. *kdnuggets.* [Online]
Available at: https://www.kdnuggets.com/2017/02/yhat-support-vector-machine.html

KHAZAN, O., 2014. *There's Something About Cities and Suicide,* The Atlantic: s.n.

quantstart, n.d. *Support Vector Machines: A Guide for Beginners,* s.l.: s.n.

Ribeiro, J. C. F. a. J. D., 2017. *Risk Factors for Suicidal Thoughts and Behaviors: A Meta-Analysis of 50,* s.l.: Psychological Bulletin.

Saxena S, F. M. C. D., 2013. *World health assembly adopts comprehensive mental health action plan ,* s.l.: s.n.

schiepek G., F. C. S. J. K. K. F. R. P. M., 2011. *Nonlinear dynamics: theoretical perspectives and application to suicidology.,* s.l.: Suicide Life Threat Behave.

Sciences., N. A. o., 2002. *ncbi,* s.l.: National Academy of Sciences..

tree, d., n.d. *geeks for geeks.* [Online]
Available at: https://www.geeksforgeeks.org/decision-tree/

Yin HL, X. L. S. Y. L. L. W. C., 2008. Relationship between suicide rate and economic growth and stock market. *dovepress.*

Appendix A: Python code Linear Regression

```
import pandas as pd

df=pd.read_csv(r'D:\thesis\scd.csv')


X = df[["country","year","sex","age","population","gdp_per_capita ($)"]]

y = df[["suicides_no"]]


from sklearn.preprocessing import LabelEncoder

lb = LabelEncoder()

X["country_code"] = lb.fit_transform(df["country"])

X["sex_code"] = lb.fit_transform(df["sex"])

X["age_code"] = lb.fit_transform(df["age"])


del X["country"]

del X["sex"]

del X["age"]

from sklearn.model_selection import train_test_split

#Train and test set

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)


#fit linear regression


lm = linear_model.LinearRegression()
```

```
model = lm.fit(X_train, y_train)

y_pred = lm.predict(X_test)

rmse = sqrt(mean_squared_error(y_test, y_pred))

print("RMSE =")

print(rmse)



#### Fit linear regression with cross validation

from sklearn.model_selection import KFold




kf = KFold(n_splits=10)

for train_index, test_index in kf.split(X,y):

    #print("TRAIN:", train_index, "TEST:", test_index)

    X_train, X_test = X.iloc[train_index], X.iloc[test_index]

    y_train, y_test = y.iloc[train_index], y.iloc[test_index]

    lr_model = lm.fit(X_train, y_train)

    y_pred = lr_model.predict(X_test)

    rmse = sqrt(mean_squared_error(y_test, y_pred))

    print("RMSE =")

    print(rmse)
```

Appendix A: Python code Decision Tree Regressor

```
import pandas as pd

from sklearn.ensemble import RandomForestRegressor

from sklearn import linear_model

from sklearn.linear_model import LinearRegression

from sklearn.preprocessing import LabelEncoder

from sklearn.model_selection import KFold

from sklearn.svm import SVR

from sklearn.model_selection import train_test_split

from math import sqrt

from sklearn.metrics import mean_squared_error

from sklearn import tree


df=pd.read_csv(r'D:\thesis\scd.csv')


X = df[["country","year","sex","age","population","gdp_per_capita ($)"]]

y = df[["suicides_no"]]


lb = LabelEncoder()

X["country_code"] = lb.fit_transform(df["country"])

X["sex_code"] = lb.fit_transform(df["sex"])

X["age_code"] = lb.fit_transform(df["age"])
```

```
del X["country"]
del X["sex"]
del X["age"]

#Train and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)

#Normal DT RMSE
clf = tree.DecisionTreeRegressor()
model = clf.fit(X_train, y_train)
y_pred = model.predict(X_test)
rmse = sqrt(mean_squared_error(y_test, y_pred))
print("RMSE =" + str(rmse))

#### Fit DT with cross validation
n_split = 10
kf = KFold(n_splits= n_split, random_state = 10 , shuffle = True)
rmse = 0
for train_index, test_index in kf.split(X,y):
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
    #print(X_train.shape)
    #print(y_train.shape )
    model = clf.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    curr_rmse = sqrt(mean_squared_error(y_test, y_pred))
    print("curr rmse = "+str(curr_rmse))
    rmse = rmse +  curr_rmse
final_rmse = rmse / n_split
print("cross validation RMSE =" + str(final_rmse))
```

Appendix A: Python code Support Vector Machines

```
# -*- coding: utf-8 -*-
"""
Created on Mon May 20 14:13:47 2019

@author: asus
"""
import pandas as pd
df=pd.read_csv(r'D:\thesis\scd.csv')

X = df[["country","year","sex","age","population","gdp_per_capita ($)"]]
y = df[["suicides_no"]]
```

```
from sklearn.preprocessing import LabelEncoder
lb = LabelEncoder()
X["country_code"] = lb.fit_transform(df["country"])
X["sex_code"] = lb.fit_transform(df["sex"])
X["age_code"] = lb.fit_transform(df["age"])

del X["country"]
del X["sex"]
del X["age"]

#Train and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)

### Fit SVM
from sklearn.model_selection import train_test_split
from sklearn.svm import SVR
svr_reg = SVR(gamma=0.001, C=1.0, epsilon=0.2)
svr_model = svr_reg.fit(X_train, y_train)
y_pred = svr_model.predict(X_test)
y_pred
rmse = sqrt(mean_squared_error(y_test, y_pred))
print("RMSE =")
print(rmse)
```

```
svr_reg = SVR(gamma=0.001, C=1.0, epsilon=0.2)

#### Fit SVM regression with cross validation

for train_index, test_index in kf.split(X,y):

    #print("TRAIN:", train_index, "TEST:", test_index)

    X_train, X_test = X.iloc[train_index], X.iloc[test_index]

    y_train, y_test = y.iloc[train_index], y.iloc[test_index]

    svr_model = svr_reg.fit(X_train, y_train)

    y_pred = svr_model.predict(X_test)

    rmse = sqrt(mean_squared_error(y_test, y_pred))

    print("RMSE =")

    print(rmse)
```

Appendix A: Python code Random Forest Regressor

```
import pandas as pd

from sklearn.ensemble import RandomForestRegressor

from sklearn import linear_model

from sklearn.linear_model import LinearRegression

from sklearn.preprocessing import LabelEncoder

from sklearn.model_selection import KFold

from sklearn.svm import SVR

from sklearn.model_selection import train_test_split
from math import sqrt
from sklearn.metrics import mean_squared_error

df=pd.read_csv(r'D:\thesis\scd.csv')

X = df[["country","year","sex","age","population","gdp_per_capita ($)"]]
y = df[["suicides_no"]]

lb = LabelEncoder()
X["country_code"] = lb.fit_transform(df["country"])
X["sex_code"] = lb.fit_transform(df["sex"])
X["age_code"] = lb.fit_transform(df["age"])

del X["country"]
del X["sex"]
del X["age"]

#Train and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
regr = RandomForestRegressor(max_depth=3, random_state=0, n_estimators=100)
model = regr.fit(X, y)

### Random FOrest
```

```
print(regr.feature_importances_)

y_pred = regr.predict(X_test)

rmse = sqrt(mean_squared_error(y_test, y_pred))

print("RMSE =")

print(rmse)



#### Fit RF with cross validation

n_split = 10

kf = KFold(n_splits= n_split, random_state = 10 , shuffle = True)

rmse = 0

for train_index, test_index in kf.split(X,y):

    #print("TRAIN:", train_index, "TEST:", test_index)

    X_train, X_test = X.iloc[train_index], X.iloc[test_index]

    y_train, y_test = y.iloc[train_index], y.iloc[test_index]

    print(X_train.shape)

    print(y_train.shape )

    RF_model = regr.fit(X_train, y_train)

    y_pred = RF_model.predict(X_test)

    rmse = rmse +  sqrt(mean_squared_error(y_test, y_pred))

final_rmse = rmse / n_split

print("cross validation RMSE =" + str(final_rmse))

regr = RandomForestRegressor(max_depth=3, random_state=0, n_estimators=100)

model = regr.fit(X, y)
```

```
### Random FOrest
print(regr.feature_importances_)
y_pred = regr.predict(X_test)
rmse = sqrt(mean_squared_error(y_test, y_pred))
print("RMSE =")
print(rmse)


#### Fit RF with cross validation
n_split = 3
kf = KFold(n_splits= n_split)
rmse = 0
for train_index, test_index in kf.split(X,y):
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
    print(X_train.shape)
    print(y_train.shape )
    RF_model = regr.fit(X_train, y_train)
    y_pred = RF_model.predict(X_test)
    rmse = rmse +  sqrt(mean_squared_error(y_test, y_pred))
final_rmse = rmse / n_split
print("cross validation RMSE =" + str(final_rmse))
```

Appendix A: Python code K-Nearest Neighbour Repressor

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn import linear_model
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import KFold
from sklearn.svm import SVR
from sklearn.model_selection import train_test_split
from math import sqrt
from sklearn.metrics import mean_squared_error
from sklearn import tree

df=pd.read_csv(r'D:\thesis\scd.csv')

X = df[["country","year","sex","age","population","gdp_per_capita ($)"]]
y = df[["suicides_no"]]

lb = LabelEncoder()
X["country_code"] = lb.fit_transform(df["country"])
X["sex_code"] = lb.fit_transform(df["sex"])
X["age_code"] = lb.fit_transform(df["age"])
```

```
del X["country"]

del X["sex"]

del X["age"]


K = 10 #Number of neighbors


#Train and test set

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)


from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler(feature_range=(0, 1))

x_train_scaled = scaler.fit_transform(X_train)

x_train = pd.DataFrame(x_train_scaled)

x_test_scaled = scaler.fit_transform(X_test)

x_test = pd.DataFrame(x_test_scaled)


from sklearn import neighbors




rmse_val = [] #to store rmse values for different k

model = neighbors.KNeighborsRegressor(n_neighbors = K)

model.fit(x_train, y_train)  #fit the model
```

```
pred=model.predict(x_test) #make prediction on test set

error = sqrt(mean_squared_error(y_test,pred)) #calculate rmse

print('RMSE value for k = ' , K , 'is:', error, 'with no cross validation')


print("Below are all cross validation values and the final CV RMSE")

#### Fit DT with cross validation

n_split = 10

kf = KFold(n_splits= n_split, random_state = 10 , shuffle = True)

rmse = 0


for train_index, test_index in kf.split(X,y):

    X_train, X_test = X.iloc[train_index], X.iloc[test_index]

    y_train, y_test = y.iloc[train_index], y.iloc[test_index]


    x_train_scaled = scaler.fit_transform(X_train)

    x_train = pd.DataFrame(x_train_scaled)

    x_test_scaled = scaler.fit_transform(X_test)

    x_test = pd.DataFrame(x_test_scaled)


    model = neighbors.KNeighborsRegressor(n_neighbors = K)

    model.fit(x_train, y_train)  #fit the model

    y_pred = model.predict(x_test) #make prediction on test set

    curr_rmse = sqrt(mean_squared_error(y_test,y_pred)) #calculate rmse
    rmse = rmse +  curr_rmse

    print('RMSE value for k= ' , K , 'is:', curr_rmse)



final_rmse = rmse / n_split

print("cross validation RMSE = " + str(final_rmse))
```