

CAUSE OF DEATH ANALYSIS

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

PARISINATI HARSHIKA (37110545)
GANIPINENI SOWMYA (37110224)



DEPARTMENT OF COMPUTER SCIENCE AND

ENGINEERING SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

MARCH 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC I 12B Status by UGC I

Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING **BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **PARISINATI HARSHIKA(37110545) and GANIPINENI SOWMYA(37110224)** who carried out the project entitled " **CAUSE OF DEATH ANALYSIS**" under my supervision from November 2020 to March 2021.

INTERNAL GUIDE

Dr. USHA NANDINI, M.E.,Ph.D

Head of the Department

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I **PARISINATI HARSHIKA** and **GANIPINENI SOWMYA** hereby declare that the Project Report entitled “ **CAUSE OF DEATH ANALYSIS**” done by us under the guidance of **Dr. USHA NANDINI, M.E.,Ph.D** Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D.**, Dean, School of Computing and **Dr.L.Lakshmanan M.E., Ph.D.**, and **Dr.S.Vigneshwari M.E., Ph.D.**, Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. USHA NANDINI, M.E.,Ph.D** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Death records can offer insights into many aspects in current existence. This venture is ready locating predominant explanation for deaths from the dataset. The fundamental cause of this project is to make consciousness approximately the deaths in everywhere in the country. So that their authorities can take important precautions for destiny and is likewise useful for coverage organization to set positive quantity for specific dying. The machine used for this project are getting to know algorithms and python programming.

This project proposes approximately predicting the reasons of dying in a specific time period through thinking about the entities with inside the statistics set. The end result offers the accuracy for all of the anticipated system getting to know fashions.

The fact that almost all of the data was categorical meant that I could not visualize the data in many different ways. I looked at this data because I was interested to see if the leading causes of death for the US, heart disease and cancer, were the same. I was also interested to see what causes of death would jump out from the data and might say something about US itself and the people who live there.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	v
	LIST OF FIGURES	viii
	LIST OF ABBREVIATIONS	ix
1	INTRODUCTION	1
	1.1 INTRODUCTION	1
	1.2 MOTIVATION	2
	1.3 SIGNIFICANCE	3
2	LITERATURE SURVEY	4
3	AIM AND SCOPE	7
	3.1 AIM	7
	3.2 OBJECTIVE	7
	3.3 SCOPE	7
	3.4 EXISTING SYSTEM	7
	3.5 PROPOSED SYSTEM	8
	3.6 SYSTEM REQUIREMENTS	8
4	SYSTEM DESIGN AND METHODOLOGY	9
	4.1 GENERAL INTRODUCTION TO MACHINE LEARNING	9
	4.2 SUPERVISED LEARNING	10
	4.3 UNSUPERVISED LEARNING	10
	4.4 SYSTEM DESIGN	11
	4.4.1 DATA COLLECTION	11
	4.4.2 DATA PRE-PROCESSING	12

	4.4.3 DATA EXTRACTION	13
	4.4.4 PREDICTION OF MAJOR CAUSE	13
	4.4.5 ALGORITHMS	14
5	RESULT AND DISCUSSION	27
	5.1 DATA PREPROCESSING	27
	5.2 PREDICTING THE TOP CAUSE	30
	5.3 REGRESSION ANALYSIS	36
6	CONCLUSION AND FUTURE WORK	41
	6.1 CONCLUSION	41
	6.2 FUTURE WORK	41
	REFERENCES	42
	APPENDIX	44
	A.SOURCE CODE	44
	B.SCREEN SHOTS	57
	C.PAPER PUBLICATION	61

LIST OF FIGURES

FIGURE No.	FIGURE NAME	PAGE No.
4.1	System Design	13
4.2	Linear Regression Model	15
4.3	Decision Tree	20
4.4	Decision Tree Architecture	21
4.5	Random Forest Architecture	24

LIST OF ABBREVIATIONS

ICD	International Classification Of Disease
DOHMH	Department of Health and Mental Hygiene
HIV	Human Immunodeficiency Viruses
AF	Atrial Fibrillation
RR	Respiratory Rate
CMHS	Center for Mental Health Services
MOR	Medical Observation Record
SMO	Sequential Minimal Optimization
CDC	Center for Disease Control's
NCHS	National Center for Health Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organization
SVM	Support Vector Machine
PCA	Principal Component Analysis

1.1 INTRODUCTION

C H A P T E R 1 I N T R O D U C T I O N

Cause of death is an official determination of conditions resulting in a human's death, which may be recorded on a death certificate. A cause of death is determined by a medical examiner. The cause of death is specific disease or injury, in contrast to the manner of death which is small number of categories like "natural", "accident", "suicide", and "homicide" which have different legal implications. International Classification of disease (ICD) codes are often used to

record manner and cause of death in a systematic way that makes it easy to compile statistics and more feasible. A study published in preventing chronic disease found that only one- third of New York city resident physicians reported believing that the present system of documentation was accurate. Half reported the inability to record “what they felt to be the correct cause of death”.

Health departments discourage listing old age as the cause of death because doing so does not benefit public health or medical research. Old age is not a scientifically recognized cause of death, there is always a more direct cause, although it may be unknown in certain cases and could be one of a number of aging- associated diseases. As an indirect or non-determinative factor, biological aging is the biggest contributor to deaths worldwide. It is estimated that of the roughly 150,000 people who die each day across the globe, about two thirds—100,000 per day—die of age-related causes. In industrialized nations the proportion is much higher, reaching 90%. Cause of death also provides information for public health planning and policy. Causes of death data can inform ideas as to what preventative measures the city should invest in to improve longevity in the city.

Machine Learning Algorithms are mostly used for predicting and analyzing from a vast dataset very easily. Predicting the future by sets of data from the admitted patient record in hospital from the years 1999–2016 are described and predicting events from extracted dataset. Different machine learning techniques will be used for predicting.

Today, analysis of mortality data increasingly uses a multiple-cause-of-death approach. Generally, cause of death statistics are derived from the so-called underlying cause of death in a process hereafter refer to as the classic method. Department of Health and Mental Hygiene (DOHMH) defines the underlying cause of death as “the disease or injury which initiated the events leading directly to death or the circumstances of the accident or violence which produced fatal injury”. However, deaths are often caused by more than one disease. More over in a world characterized by an ageing population and decreasing mortality and fertility. Death due to infectious disease is progressively being replaced by death due to chronic and degenerative diseases.

1.2 MOTIVATION

Many of the scientific discussions and studies in biomedical and healthcare domains address tasks whose end goal is to prevent death or diseases. Since the emergence of the big data science, numerous machine learning based techniques and technologies have been proposed and applied to improve human health by solving different computational challenges that we face today. In fact it is widely believed that counting the dead is a significant investment to reduce the premature mortality. There have been a number of studies that have proven to offer profound impacts on our understanding of the major causes of death using the statistical analysis of recorded death data. In light of these studies, we were interested to investigate the feasibility of this emerging field in learning hidden complex patterns that are available in the haystack of mortality datasets. causes of death data provide a valuable source of information that have been used to analyse the death event in chronic diseases such as the HIV and the lung disease, to identify problems with the process of coding/recording cause-of-death information. Moreover, these data can be potentially used in analysis of disease diffusion for controlling plagues and other epidemics and may provide a better understanding of multi-morbid associations between conditions leading to death. As such, designing advanced analytics pipelines for discovering descriptive statistics and trajectories is highly crucial. In particular,

the multiple-cause-of- death data are unstructured and are often inaccurate and noisy.

Moreover, the high number of ICD-9/10 mortality codes makes analysis of multiple-cause associations even more challenging. These altogether, call for advanced techniques for mining large datasets of unstructured, high dimensional, and noisy structure.

Despite the importance of the subject, only a handful of studies have so far conducted research seeking to relate multiple causes of death to each other or to other factors. These studies are often limited to classical statistical methods (measures) that do not scale up efficiently and are put into four major categories:

- 1) Univariate measures, consisting of counts and frequencies
- 2) Cross-tabular measures, which incorporate variables that identify the roles (e.g. contributory, non-contributory, complication and underlying) associated with multiple death causes
- 3) Measures of association, in which some measure of multiple mentions of a cause is related to some measure of mentions of the underlying cause
- 4) Derived measures, where univariate measures such as multiple-cause rates are integrated to build higher order

1.3 SIGNIFICANCE

The main aim of this proposed system is to analyze the top most causes of death from the dataset by using different machine learning algorithms. This analysis is very helpful for the insurance companies to set the policy price and it is also for the governments to take preventive measures to reduce the deaths. In some hospitals deaths are happening due to lack of equipment and necessary medicine. By using this analysis government can provide necessary equipment and medicine for the hospitals to reduce deaths. This analysis is used for the government to take necessary actions on homicide deaths and also used for reducing the accidental deaths by providing causing awareness in the people. The proposed system provides the analysis of deaths caused due to different diseases like heart attack, cancer, accidents, etc., using different machine learning algorithms like linear

regression, random forest method, decision tree and support vector machine in different states. We are going to find out which algorithm is most efficient.

CHAPTER 2

LITERATURE SURVEY

TRENDS IN THE LEADING CAUSES OF DEATH IN THE UNITED STATES, 1970-2002

AUTHOR: Ahmedin Jemal, DVM, PhD; Elizabeth Ward, PhD; Yongping Hao, PhD

PUBLISHED YEAR: September 14, 2005

The absolute number of deaths and age at death continue to increase in the United States. These temporal trends have major implications for health care and health care costs in an aging population. Age-standardized death rates from all causes have decreased in the United States since the 1960s. Understanding these trends and the relationship between the age-standardized death rates and the actual number of deaths that occur can provide valuable insight into the forces that shape the nation's health. They examined trends in death rates and number of deaths from the 6 leading causes in the United States and considered the relationship of these trends to disease prevention and health care in an aging population.

CAUSES OF DEATH AND INFLUENCING FACTORS IN PATIENTS WITH ATRIAL FIBRILLATION

A Competing-Risk Analysis from the Randomized Evaluation of Long-Term Anticoagulant Therapy Study

AUTHORS: Jean-Yves Le Heuzey, Stuart Connolly, Sean Yang, Janice Pogue, Martina Brueckmann, John Eikelboom, Ellison Themeles, Michael Ezekowitz, Lars Wallentin, and Salim Yusuf

PUBLISHED YEAR: 9 Sep 2013

The majority of deaths are not related to stroke in a contemporary anticoagulated AF population. Dabigatran reduces deaths resulting from vascular causes (but not other causes) compared with warfarin in patients with AF. These results emphasize the

need to identify interventions beyond effective anticoagulation to further reduce mortality in

atrial fibrillation, especially targeting cardiac mortality.

UNDERNUTRITION AS AN UNDERLYING CAUSE OF CHILD DEATHS ASSOCIATED WITH DIARRHEA, PNEUMONIA, MALARIA, AND MEASLES

AUTHORS: *Laura E Caulfield, Mercedes de Onis, Monika Blössner, Robert E Black*

PUBLISHED YEAR: July 2004

The objectives were to examine whether the risk of dying because of underweight varies by cause of death and to estimate the fraction of deaths by cause attributable to underweight. The RR of mortality because of low weight-for-age was elevated for each cause of death and for all-cause mortality.

Overall, 52.5% of all deaths in young children were attributable to undernutrition, varying from 44.8% for deaths because of measles to 60.7% for deaths because of diarrhea.

MORTALITY AND CAUSE OF DEATH AMONG PSYCHIATRIC PATIENTS: A 20-YEAR CASE-REGISTER STUDY IN AN AREA WITH A COMMUNITY-BASED SYSTEM OF CARE

AUTHORS: *L. Grigoletti¹ *, G. Perini¹ , A. Rossi¹ , A. Biggeri² , C. Barbui¹ , M. Tansella¹ and F. Amaddeo¹*

PUBLISHED YEAR: 2009

This study found that psychiatric patients in contact with a CMHS have an almost twofold higher mortality rate than the general population. These findings demonstrate that, since the closure of long-stay psychiatric hospitals, the physical health care of people with mental health problems is often neglected and clearly requires greater attention by health-care policymakers, services and professionals.

SERUM URIC ACID AND CARDIOVASCULAR MORTALITY THE NHANES I EPIDEMIOLOGIC FOLLOW-UP STUDY, 1971-1992

AUTHORS: *Jing Fang, MD; Michael H. Alderman, MD*

PUBLISHED YEAR: May 10, 2000

The NHANES I, sponsored by the National Center for Health Statistics, was conducted from 1971 to 1975 on a nationwide probability sample of the US civilian noninstitutionalized population. The survey included a standardized medical examination and questionnaires that addressed a variety of health-related topics. Low-income persons, the elderly, and women of childbearing age were oversampled. This updated report extends the earlier study of the association of serum uric acid levels with cardiovascular mortality and ischemic heart disease mortality in men and women.

DEATH PREDICTION AND ANALYSIS USING WEB MINING TECHNIQUES

AUTHORS: Hesham Abdo Ahmed Aqlan ,Shoiab Ahmed Ajit Danti

PUBLISHED YEAR: 2017

In this paper, death prediction and Analysis using web mining techniques is proposed. In the proposed approach four methods have been employed for prediction of death from the large-scale digital histories of 18 years duration captured from news reports of Queensland Government archive to make realtime predictions about the likelihoods of future death and disease events of interest. Based on the evaluation of results and graphical analysis reveal interesting facts that SMOReg and Linear Regression algorithms are better suited for forecasting future causes of death related information.

3.1 AIM

C H A P T E R 3 A I M A N D S C O P E

The main aim of the proposed system is to find the major cause of death among the entire dataset of years 1999-2017. And also wanted to find the best Machine Learning model which can give better accuracy in predicting the major cause of death. This analysis is majorly helpful and useful for the government to take necessary precautions for the future. And it can also be helpful for the insurance companies so that they can fix a particular price for the particular cause of death.

3.2 OBJECTIVE

The main objective of this project is to find the major causes of death in particular year from the dataset. It is useful for the insurance companies to make informed decisions on what prices to set for different kinds of insurance and coverage. And also helpful to give awareness for the government so that they can take required precautions in future. Today analysis of mortality data mostly using multiple cause of death approach. Generally cause of death statistics are derived from the underlying cause of death using classic method.

3.3 SCOPE

Cause of death analysis provides information for public health planning and policy. Cause of death data can inform ideas as to what preventative measures the city should invest in to improve longevity in the city.

3.4 EXISTING SYSTEM

It is very much important to know the cause of death for people. Knowing the cause of death would be a great help for health and life insurance companies to know so that they can make informed decisions on what prices to set for different kinds of insurance and coverage. In olden days, to realize the reason of dying, we need to personally go to the door-to-door survey for enquiry and need to make a list of deaths which is a time-consuming process.

But now we can store the information once they get admitted in any hospital which is then recorded in the form of database. From the base paper and other research papers observed we get to realize the unique approaches of finding the causes of death.

3.5 PROPOSED SYSTEM

In this proposed approach, statistics is extracted for prediction of death event and later preprocessed and differentiated numerous machine learning algorithms on reasons of dying dataset of USA. Then diagnosed the satisfactory overall performance set of rules for prediction and forecasting of reason for dying illness. And also predicted the major cause of deaths for a particular country from the data set which we got from an open source website by using python programming along with Machine Learning algorithms and also differentiated each entity using bar graphs.

3.6 SYSTEM REQUIREMENTS

Hardware Requirements:

- System: Pentium IV 2.4 GHz
- Hard Disk: 40GB
- Floppy Drive: 1.44Mb
- RAM : 512Mb

Software Requirements:

- Operating System : Windows 7 Ultimate
- Coding Language : Python
- Front-End : Python
- Designing : HTML, CSS, JavaScript
- Database : MySQL

CHAPTER 4

SYSTEM DESIGN AND METHODOLOGY

4.1 GENERAL INTRODUCTION TO MACHINE LEARNING

Machine learning is the act of giving computers the ability to learn without explicitly programming them. This is done by giving data to computers and having them transform the data into decision models which are then used for future predictions. Machine learning is a type of technology that aims to learn from experience. For example, as a human, you can learn how to play chess simply by observing other people playing chess. In the same way, computers are programmed by providing them with data from which they learn and are then able to predict future elements or conditions. Let's say, for instance, that you want to write a program that can tell whether a certain type of fruit is an orange or a lemon. You might find it easy to write such a program and it will give the required results, but you might also find that the program doesn't work effectively for large datasets. This is where machine learning comes into play.

There are various steps involved in machine learning:

1. collection of data
2. filtering of data
3. analysis of data
4. algorithm training
5. testing of the algorithm
6. using the algorithm for future predictions

Machine learning uses different kinds of algorithms to find patterns, and these algorithms are classified into two groups:

- supervised learning
- unsupervised learning

4.2 SUPERVISED LEARNING

Supervised learning is the science of training a computer to recognize elements by giving it sample data. The computer then learns from it and is able to predict future datasets based on the learned data. For example, you can train a computer to filter out spam messages based on past information.

Supervised learning has been used in many applications, e.g. Facebook, to search images based on a certain description. You can now search images on Facebook with words that describe the contents of the photo. Since the social networking site already has a database of captioned images, it is able to search and match the description to features from photos with some degree of accuracy.

There are only two steps involved in supervised learning:

- training
- testing

Some of the supervised learning algorithms include:

- decision trees
- support vector machines
- naive Bayes
- k-nearest neighbour
- linear regression

4.3 UNSUPERVISED LEARNING

Unsupervised learning is when you train your machine with only a set of inputs. The machine will then be able to find a relationship between the input data and any other you might want to predict. Unlike in supervised learning, where you present a machine with some data to train on, unsupervised learning is meant to make the

computer find patterns or relationships between different datasets.

Unsupervised learning can be further subdivided into:

- clustering
- association

Clustering: Clustering means grouping data inherently. For example, you can classify the shopping habits of consumers and use it for advertising by targeting the consumers based on their purchases and shopping habits.

Association: Association is where you identify rules that describe large sets of your data. This type of learning can be applicable in associating books based on author or category, whether motivational, fictional, or educational books.

Some of the popular unsupervised learning algorithms include:

- k-means clustering
- hierarchical clustering

Unsupervised learning will be an important technology in the near future. This is due to the fact that there is a lot of unfiltered data which has not yet been digitized.

4.4 SYSTEM DESIGN

4.4.1 Data Collection

The statistics we used for this project is from the Center for Disease Control's (CDC) National Center for Health Statistics (NCHS). They preserve the database with age-wise death rates and counts throughout the united states .The statistics is grouped by cause, year, state from the years 1999-2016. Reason of dying records is derived from the so known as underlying reason of dying in a manner hereafter refed to because the traditional approach. Department of Health and Mental Hygiene

(DOHMH) defines the underlying reason of dying because

“the ailment or harm which initiated the occasions main at once to dying or the situations of the twist of a fate or violence which produced deadly effects.

4.4.2 Data Pre-processing

Data preprocessing is a manner of preparing the raw data and making it appropriate for a machine learning model. Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we preprocess our data before feeding it into our model. Data Processing is a task of converting data from a given form to a much more usable and desired form i.e. making it more meaningful and informative. Using Machine Learning algorithms, mathematical modelling and statistical knowledge, this entire process can be automated.

Data preprocessing is required obligations for cleaning the data and making it appropriate for a machine learning model which additionally increases the accuracy and performance of a machine learning model.

The output of this complete process can be in any desired form like graphs, videos, charts, tables, images and many more, depending on the task we are performing and the requirements of the machine. This might seem to be simple but when it comes to really big organizations like Twitter, Facebook, Administrative bodies like Paliament, UNESCO and health sector organizations, this entire process needs to be performed in a very structured manner.

In this the data will be read from the csv files which are taken from the Center for Disease Control's (CDC) National Center for Health Statistics (NCHS) and are uploaded. Then the preprocessing of the dataset is done in which data gets cleaned and even removes unwanted columns if required.

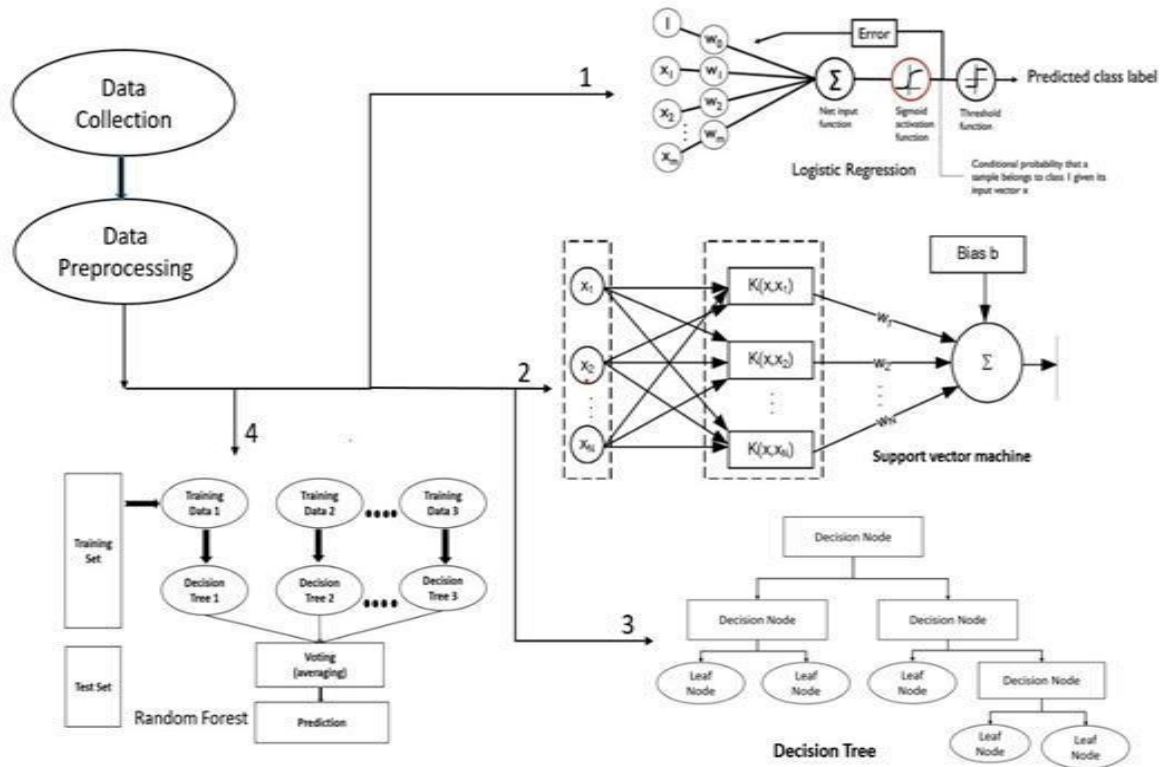


Fig.4.1 : System Design

4.4.3 Data Extraction

In this the data is transformed and gets ready for the prediction. Then we will consider few factors and find the major cause of death in USA. They are

- The main leading cause of dying in the USA
- The current trend for the main leading cause of death
- Which factors have minimum impact on these deaths

4.4.4 Prediction of major cause

Firstly, the data is grouped by top causes of death .From that result we get to know that the cause of death is “Heart disease” and a little difference of “Cancer”. So, we performed Hypothesis Testing to find the difference in death rates between Heart disease and Cancer. Then we created data frames for comparing the both causes and found that heart disease is causing majorly ahead of cancer.

The very next thing we wanted to know is the major factors causing the death by heart disease. So, we considered few factors such as

- Population Density
- Age Demographic
- Gender Demographic
- Alcohol Consumption
- Smoking Population
- Diabetes Rates

For each factor the data is scraped and converted to csv file and then the process is performed. Then we performed regression analysis for all the factors which are

4.4.5 Algorithms

Algorithms used in cause of death analysis:

- Linear Regression Model
- Support vector machine
- Decision Tree Model
- Random Forest Mode
- Logistic regression Model

1. Linear Regression

Linear Regression is a supervised machine learning algorithm. It is a statistical approach that is used for predictive evaluation. Linear regression algorithm shows a linear relationship among a dependent (y) and one or more independent (x) variables with a sloped straight line, consequently known as linear regression. It observes how the value of the dependent variable is changing according to the value of the independent variable.

The equation has the form $Y = a + Bx$, where Y is the dependent variable (that is the variable which goes on the Y axis), X is the independent variable (it is plotted on the X axis), b is the slope of the line and a is the y -intercept.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

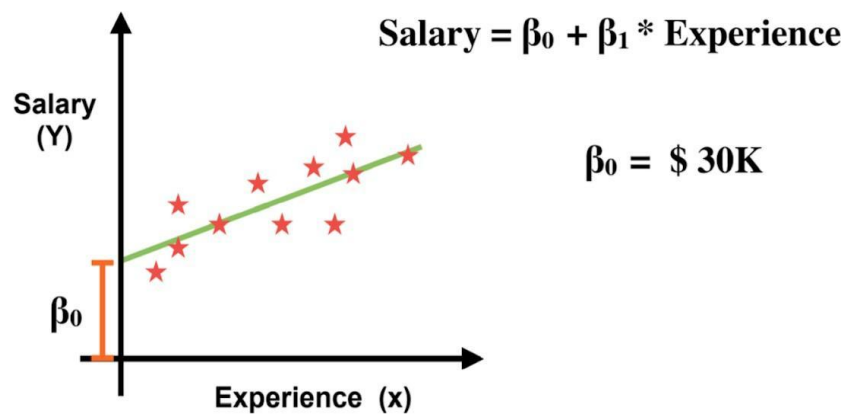


Fig. 4.2 : Linear Regression Model

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be

used to make a prediction of the response.

If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

2. Logistic Regression

It is a form of regression that lets in the prediction of discrete variables through a combination of continuous & discrete predictors.

In logistic regression, the final results variable is binary and the purpose of the evaluation is to assess the effects of multiple explanatory variables which may be and/or categorical.

The PCA does an unsupervised dimensionality reduction, while the logistic regression does the prediction. We use a GridSearchCV to set the dimensionality of the PCA.

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

is the equation used for logistic regression. In the result we get the accuracy as 0.80642373817349.

We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function.

The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

DIFFERENCE BETWEEN LINEAR AND LOGISTIC REGRESSION

The linear regression models data using continuous numeric values.

Linear regression requires to establish the linear relationship among dependent and independent variables whereas it is not necessary for logistic regression.

In the linear regression, the independent variable can be correlated with each other. In logistic regression, the variable must not be correlated with each other.

3.Support Vector Machine

Support Vector Machine is a supervised learning model associated with learning algorithms that analyse data and recognise patterns used for regression analysis. Support vector machine is extremely favoured by many as it produces notable correctness with less computation power. It is mostly used in classification problems. The intention of the support vector machine algorithm is to find a hyper plane in an N-dimensional space that separately classifies the data points. In 1960s, SVMs were first introduced but later they got refined in 1990.

SVMs have their unique way of implementation as compared to other machine learning algorithms.

Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables. I have combined SVM, PCA, and Grid-search to create a pipeline to find best parameters for binary classification & eventually plot a decision boundary to present how good our algorithm has performed. As the result it achieved a 90% accuracy with all cross-sections of test data.

In the SVM classifier, it is easy to have a linear hyper-plane between these two classes. But, another burning question which arises is, should we need to add this feature manually to have a hyper-plane. No, the SVM algorithm has a technique called the **kernel trick**. The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem

It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs you've defined. In Python, scikit-learn is a widely used library for implementing machine learning algorithms. SVM is also available in the scikit-learn library and we follow the same structure for using it (Import library, object creation, fitting model and prediction).

DIFFERENCE BETWEEN SVM AND LOGISTIC REGRESSION

SVM tries to find the “best” margin (distance between the line and the support vectors) that separates the classes and this reduces the risk of error on the data, while logistic regression does not, instead it can have different decision boundaries with different weights that are near the optimal point.

SVM is based on geometrical properties of the data while logistic regression is based on statistical approaches.

The risk of overfitting is less in SVM, while Logistic regression is vulnerable to overfitting.

SVM works well with unstructured and semi-structured data like text and images while logistic regression works with already identified independent variables.

Logistic regression and SVM are great tools for training classification and regression problems. It is good to know when to use either of them so as to save computational cost and time.

3. Decision Tree

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which

expands on further branch and constructs a tree-like structure.

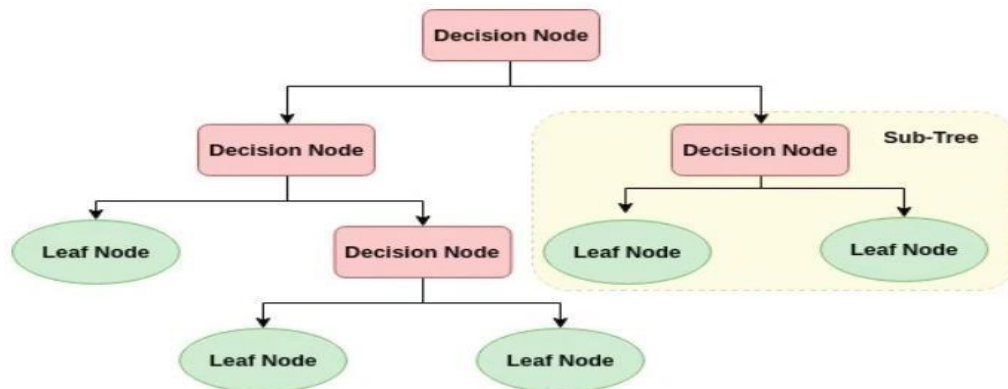


Fig . 4.3 : Decision Tree

A decision tree is a decision support tool that uses a tree-like model of choice and their feasible consequences, including chance event results, resource costs, and utility. It is one manner to display an algorithm that most effectively contains conditional control statements. In this the data is spilt into few parameters. In this information gain is used and then Data gets divided into separate rows. As the result we get the best decision tree pipeline score accuracy of 69%.

How does the Decision Tree algorithm work?

The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures (ASM) to split the records.
2. Make that attribute a decision node and breaks the dataset into smaller subsets.
3. Starts tree building by repeating this process recursively for each child until one of the conditions will match:

- All the tuples belong to the same attribute value.

- There are no more remaining attributes.
- There are no more instances.

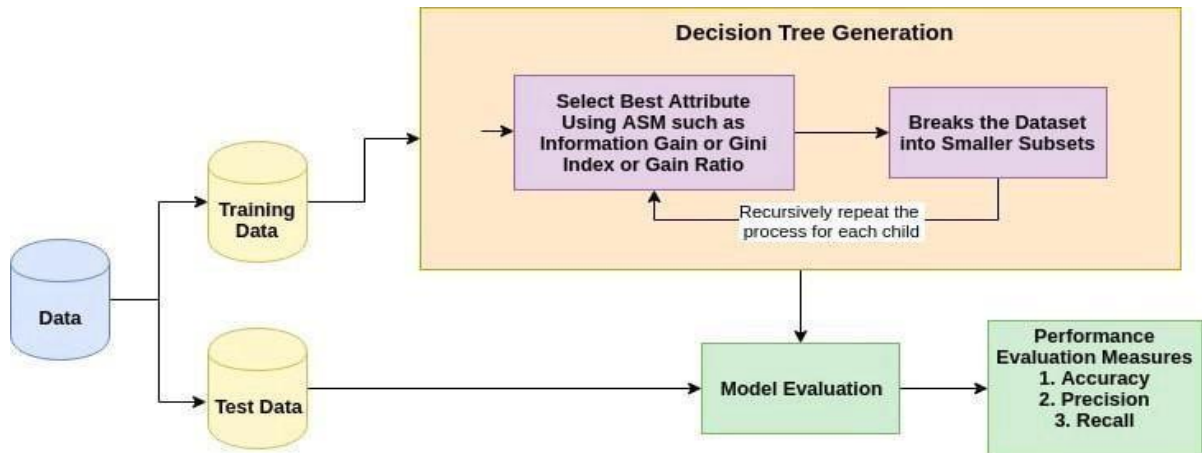


Fig .4.4 : Decision Tree Architecture

ADVANTAGES OF DECISION TREES

• COMPREHENSIVE

A significant advantage of a decision tree is that it forces the consideration of all possible outcomes of a decision and traces each path to a conclusion. It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis.

• Specific

Decision trees assign specific values to each problem, decision path and outcome. Using monetary values makes costs and benefits explicit. This approach identifies the relevant decision paths, reduces uncertainty, clears up ambiguity and clarifies the financial consequences of various courses of action.

When factual information is not available, decision trees use probabilities for conditions to keep choices in perspective with each other for easy comparisons.

- **Easy to Use**

Decision trees are easy to use and explain with simple math, no complex formulas. They present visually all of the decision alternatives for quick comparisons in a format that is easy to understand with only brief explanations.

They are intuitive and follow the same pattern of thinking that humans use when making decisions.

- **Versatile**

A multitude of business problems can be analyzed and solved by decision trees. They are useful tools for business managers, technicians, engineers, medical staff and anyone else who has to make decisions under uncertain conditions.

The algorithm of a decision tree can be integrated with other management analysis tools such as Net Present Value and Project Evaluation Review Technique (PERT).

Simple decision trees can be manually constructed or used with computer programs for more complicated diagrams.

Decision trees are a common-sense technique to find the best solutions to problems with uncertainty. Should you take an umbrella to work today? To find out, construct a simple decision-tree diagram.

3.Random Forest Method

Random Forest is a popular machine learning algorithm that belongs to the

supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble

learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, “Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.” Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

In this method it selects some random samples from the data and constructs decision trees for each sample and gets a result from each decision tree .Then it performs voting and predicts the final result .It can be used for both classification and also for regression problems. If there are a greater number of trees in forest then the accuracy will be highest. It considers the highest voted target as the final prediction. In this we got the best random forest pipeline score with accuracy of 77%.

Random forests have a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample.

In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to

the other non-linear classification algorithms.

How does the algorithm work?

It works in four steps:

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction

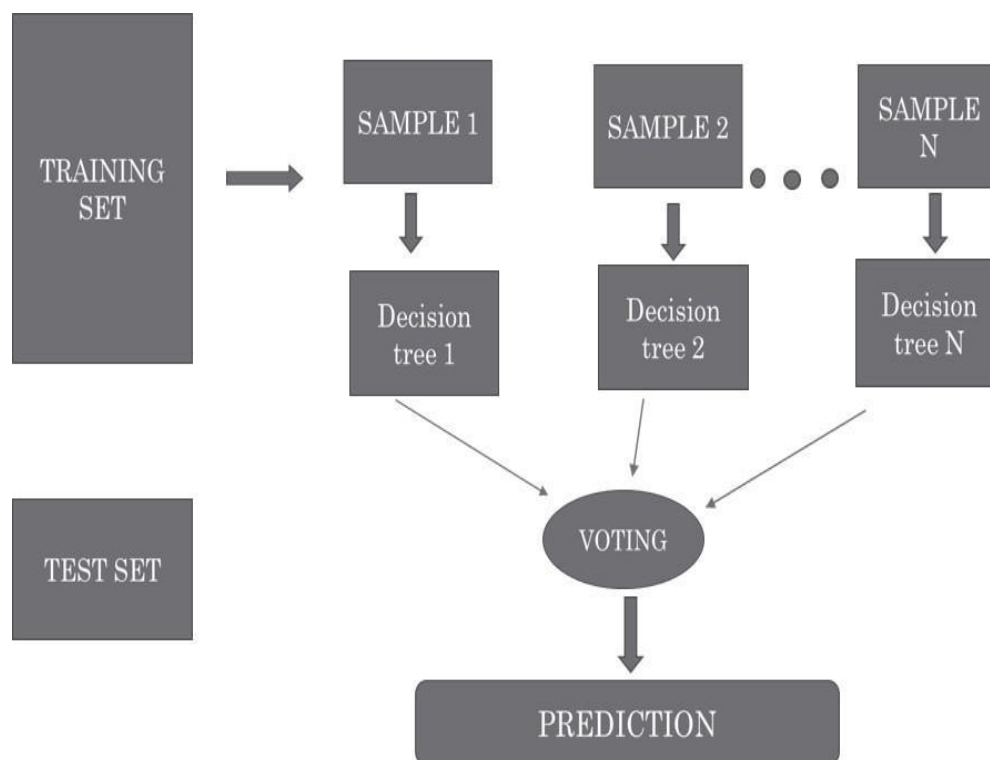


Fig.4.5 : Random Forest Architecture

Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it

searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

ADVANTAGES AND DISADVANTAGES OF THE RANDOM FOREST ALGORITHM

One of the biggest advantages of random forest is its versatility. It can be used for both regression and classification tasks, and it's also easy to view the relative importance it assigns to the input features.

Random forest is also a very handy algorithm because the default hyperparameters it uses often produce a good prediction result. Understanding the hyperparameters is pretty straightforward, and there's also not that many of them.

One of the biggest problems in machine learning is overfitting, but most of the time this won't happen thanks to the random forest classifier. If there are enough trees in the forest, the classifier won't overfit the model.

The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model.

And, of course, random forest is a predictive modeling tool and not a descriptive tool, meaning if you're looking for a description of the relationships in your data, other approaches would be better.

DIFFERENCE BETWEEN DECISION TREES AND RANDOM FORESTS

While random forest is a collection of decision trees, there are some differences.

If you input a training dataset with features and labels into a decision tree, it will formulate some set of rules, which will be used to make the predictions.

For example, to predict whether a person will click on an online advertisement, you might collect the ads the person clicked on in the past and some features that describe his/her decision. If you put the features and labels into a decision tree, it will generate some rules that help predict whether the advertisement will be clicked or not. In comparison, the random forest algorithm randomly selects observations and features to build several decision trees and then averages the results.

Another difference is "deep" decision trees might suffer from overfitting. Most of the time, random forest prevents this by creating random subsets of the features and building smaller trees using those subsets. Afterwards, it combines the subtrees. It's important to note this doesn't work every time and it also makes the computation slower, depending on how many trees the random forest builds.

CHAPTER 5

RESULT & DISCUSSION

5.1 DATA PREPROCESSING

Modelling data using machine learning techniques . The process of using a machine to predict future events based on known past events .The data is used for predicting the past analysis which may be helpful for the future. In data processing 17 year of data set has been considered. From this the maximum causes of death for all diseases have been computed. And also found the main cause of major deaths i.e., Heart disease. Through which we found that males of age above 70 are affected the most.

Then we proposed machine learning algorithms for finding the best way to predict the heart disease. Such as using Decision Tree algorithm, Random Forest method, Support Vector Machine and Logistic Regression. The results of all the algorithms have been evaluated and found that Support vector machine algorithms got the highest accuracy. So, the predictions says that it is the best method to find the heart disease death analysis.

The following screenshot is the data which we have collected :

Out[2]:

	Unnamed: 0	year	_113_cause_name	cause_name	state	deaths	aadr
0	0	2012	Nephritis, nephrotic syndrome and nephrosis (N...	Kidney disease	Vermont	21	2.6
1	1	2016	Nephritis, nephrotic syndrome and nephrosis (N...	Kidney disease	Vermont	30	3.7
2	2	2013	Nephritis, nephrotic syndrome and nephrosis (N...	Kidney disease	Vermont	30	3.8
3	3	2000	Intentional self-harm (suicide) (*U03,X60-X84,...	Suicide	District of Columbia	23	3.8
4	4	2014	Nephritis, nephrotic syndrome and nephrosis (N...	Kidney disease	Arizona	325	4.1

Fig 5.1 : Sample Data

The very next step is data pre-processing in which the initial cleaning is done before joining the main data. In this some of the columns are deleted for the better performance of the code. The output is as follows:

Out[6]:

	Year	State	Age Group	Male	Population
0	1999	Alabama	< 1 year	False	28712
1	1999	Alabama	< 1 year	True	30075
2	1999	Alabama	1-4 years	False	115715
3	1999	Alabama	1-4 years	True	120438
4	1999	Alabama	5-9 years	False	153779

Fig.5.2 DATA AFTER PREPROCESSING

Now the data gets joined in the main dataset. In addition to population by state, gender and age percentages should also be joined with primary dataset.

1. Aggregate Total Population by State & Year
2. Aggregate Male Population by State & Year
3. Aggregate Age Groups by State & Year

Out[8]:

	Year	Cause_Name	State	Deaths	Population
0	2016	Unintentional injuries	Alabama	2755	4860545
1	2016	Alzheimer's disease	Alabama	2507	4860545
2	2016	Cancer	Alabama	10419	4860545
3	2016	CLRD	Alabama	3326	4860545
4	2016	Diabetes	Alabama	1183	4860545

Fig.5.3 TOTAL POPULATION BY STATE & YEAR

Out[9]:

	Year	Cause_Name	State	Deaths	Population_Total	Population_Male
2	2016	Cancer	Alabama	10419	4860545	0.484347
12	2016	Cancer	Alaska	995	741522	0.523475
22	2016	Cancer	Arizona	11876	6908642	0.497096
32	2016	Cancer	Arkansas	6612	2988231	0.491298
42	2016	Cancer	California	59515	39296476	0.496812

Fig.5.4 MALE POPULATION BY STATE & YEAR

Out[10]:

	Year	Cause_Name	State	Deaths	Population_Total	Population_Male	Population_<1 year	Population_1-4 years	Population_5-9 years	Population_10-14 years	...	Population_40-44 years
0	2016	Unintentional injuries	Alabama	2755	4860545	0.484347	0.012039	0.048362	0.062506	0.063089	...	0.06043
1	2016	Alzheimer's disease	Alabama	2507	4860545	0.484347	0.012039	0.048362	0.062506	0.063089	...	0.06043
2	2016	Cancer	Alabama	10419	4860545	0.484347	0.012039	0.048362	0.062506	0.063089	...	0.06043
3	2016	CLRD	Alabama	3326	4860545	0.484347	0.012039	0.048362	0.062506	0.063089	...	0.06043
4	2016	Diabetes	Alabama	1183	4860545	0.484347	0.012039	0.048362	0.062506	0.063089	...	0.06043

Fig.5.5 AGE GROUP BY STATE & YEAR

5.2 PREDICTING TOP CAUSE

Next is finding the top cause of death from the data by grouping them.

Out[14]:

	Deaths	Population	Deaths_per_Capita
Cause_Name			
Alzheimer's disease	1373412	5436981948	0.000253
CLRD	2434726	5436981948	0.000448
Cancer	10244536	5436981948	0.001884
Diabetes	1316379	5436981948	0.000242
Heart disease	11575183	5436981948	0.002129
Influenza and pneumonia	1038969	5436981948	0.000191
Kidney disease	807980	5436981948	0.000149
Stroke	2580140	5436981948	0.000475
Suicide	649843	5436981948	0.000120
Unintentional injuries	2177884	5436981948	0.000401

Fig 5.6 FINDING TOP CAUSE OF DEATH

From the above output we get to know that heart disease and cancer are with slight difference .So we did comparision between heart disease and cancer.

Out[22]:

	Deaths_per_Capita
Year	
1999	0.002599
2000	0.002519
2001	0.002457
2002	0.002423
2003	0.002361
2004	0.002228
2005	0.002207
2006	0.002117
2007	0.002045
2008	0.002028
2009	0.001954
2010	0.001932
2011	0.001914
2012	0.001910
2013	0.001932
2014	0.001928
2015	0.001974
2016	0.001964

Fig 5.7 COMPARING CANCER AND HEART DISEASE

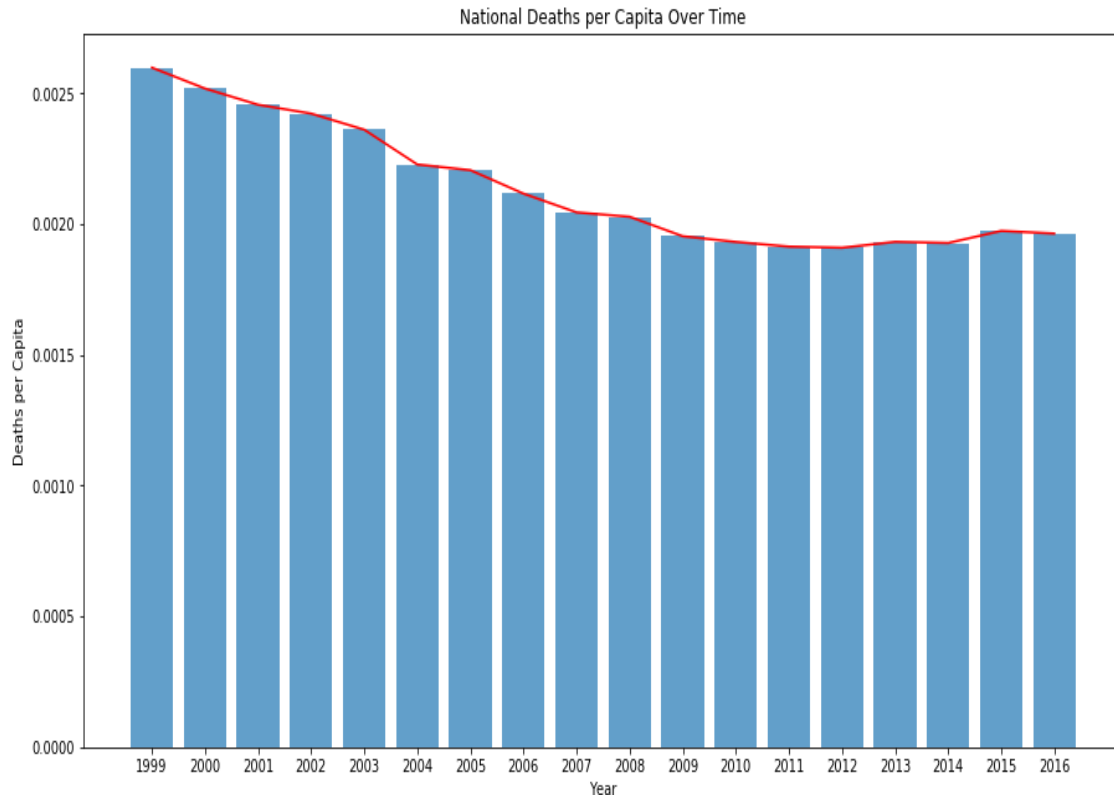


Fig 5.7 COMPARING HEART DISEASE FROM 1999-2016

Then next finding the factors which has greatest impact of heart disease .The factors include:

- Population Density
- Age Demographic
- Gender Demographic
- Alcohol Consumption
- Smoking Population
- Diabetes Rates

Out[33]:

	Year	Cause_Name	State	Deaths	Population	Population_Male	Population_<_1_year	Population_1_4_years	Population_5_9_years	Population_10_14
0	2016	Heart disease	Alabama	12832	4860545	0.484347	0.012039	0.048362	0.062506	0.063089
1	2015	Heart disease	Alabama	12981	4850858	0.484665	0.012179	0.048409	0.062936	0.063828
2	2014	Heart disease	Alabama	12461	4840037	0.484787	0.012057	0.048594	0.063281	0.064846
3	2013	Heart disease	Alabama	12472	4827660	0.484981	0.011886	0.049125	0.063570	0.065616
4	2012	Heart disease	Alabama	12036	4813946	0.485060	0.012156	0.049996	0.063502	0.066419

5 rows x 27 columns

Fig 5.8 POPULATION DENSITY OUTPUT

Out[38]:

	Year	beer	wine	spirits	all_bevs
count	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000
mean	1996.500000	1.287589	0.341455	0.845817	2.475192
std	11.545974	0.226351	0.179546	0.349340	0.613162
min	1977.000000	0.630000	0.080000	0.360000	1.190000
25%	1986.750000	1.130000	0.210000	0.640000	2.120000
50%	1996.500000	1.280000	0.310000	0.765000	2.360000
75%	2006.250000	1.410000	0.450000	0.950000	2.700000
max	2016.000000	2.170000	1.190000	3.560000	6.320000

Fig 5.9 FINDING TYPE OF ALCOHOL EFFECTING HEART DISEASE

Out[40]:

	Year	Cause_Name	State	Deaths	Population	Population_Male	Population_<_1_year	Population_1_4_years	Population_5_9_years	Population_10_14
0	2016	Heart disease	Alabama	12832	4860545	0.484347	0.012039	0.048362	0.062506	0.063089
1	2015	Heart disease	Alabama	12981	4850858	0.484665	0.012179	0.048409	0.062936	0.063828
2	2014	Heart disease	Alabama	12461	4840037	0.484787	0.012057	0.048594	0.063281	0.064846
3	2013	Heart disease	Alabama	12472	4827660	0.484981	0.011886	0.049125	0.063570	0.065616
4	2012	Heart disease	Alabama	12036	4813946	0.485060	0.012156	0.049996	0.063502	0.066419

5 rows x 31 columns



Fig 5.10 TABLE OF ALCOHOL EFFECTED POPULATION

Out[41]:

	Year	State	Value
0	1999	Alabama	24.6
1	1999	Alaska	26.1
2	1999	Arizona	21.8
3	1999	Arkansas	25.9
4	1999	California	19.2

Fig 5.11 STATES EFFECTED BY SMOKING IN 1999

Out[43]:

	Year	Cause_Name	State	Deaths	Population	Population_Male	Population_<_1_year	Population_1_4_years	Population_5_9_years	Population_10_14
576	2016	Heart disease	New York	44076	19836286	0.485390	0.011933	0.047238	0.057361	0.057835
577	2015	Heart disease	New York	44450	19819347	0.485129	0.012172	0.047621	0.057330	0.058471
578	2014	Heart disease	New York	43116	19773580	0.484966	0.012078	0.047811	0.057649	0.059268
579	2013	Heart disease	New York	44039	19712514	0.484863	0.012248	0.047577	0.058414	0.059863
580	2012	Heart disease	New York	43795	19625409	0.484642	0.012302	0.047582	0.058847	0.060557
581	2011	Heart disease	New York	44501	19526372	0.484283	0.012515	0.047485	0.059219	0.061432
582	2010	Heart disease	New York	44981	19405185	0.483988	0.012160	0.047615	0.059873	0.062281
583	2009	Heart disease	New York	47283	19307066	0.483784	0.012274	0.047670	0.060170	0.063315
584	2008	Heart disease	New York	49324	19212436	0.483728	0.012593	0.047745	0.060115	0.064581
585	2007	Heart disease	New York	49528	19132335	0.483518	0.012640	0.047834	0.060363	0.066163
586	2006	Heart disease	New York	50470	19104631	0.483298	0.012353	0.048417	0.060957	0.067913

Fig 5.12 DEATHS IN NEWYORK CITY BY SMOKING

Out[48]:

	Year	State	Diabetes
4	2016	Alabama	1183
14	2016	Alaska	124
24	2016	Arizona	2025
34	2016	Arkansas	920
44	2016	California	9124

Fig 5.13 DEATHS CAUSED BY DIABETES

Out[49]:

	Year	Cause_Name	State	Deaths	Population	Population_Male	Population_<_1_year	Population_1_4_years	Population_5_9_years	Population_10_14
0	2016	Heart disease	Alabama	12832	4860545	0.484347	0.012039	0.048362	0.062506	0.063089
1	2015	Heart disease	Alabama	12981	4850858	0.484665	0.012179	0.048409	0.062936	0.063828
2	2014	Heart disease	Alabama	12461	4840037	0.484787	0.012057	0.048594	0.063281	0.064846
3	2013	Heart disease	Alabama	12472	4827660	0.484981	0.011886	0.049125	0.063570	0.065616
4	2012	Heart disease	Alabama	12036	4813946	0.485060	0.012156	0.049996	0.063502	0.066419

5 rows x 33 columns

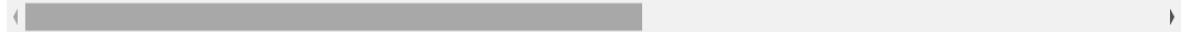
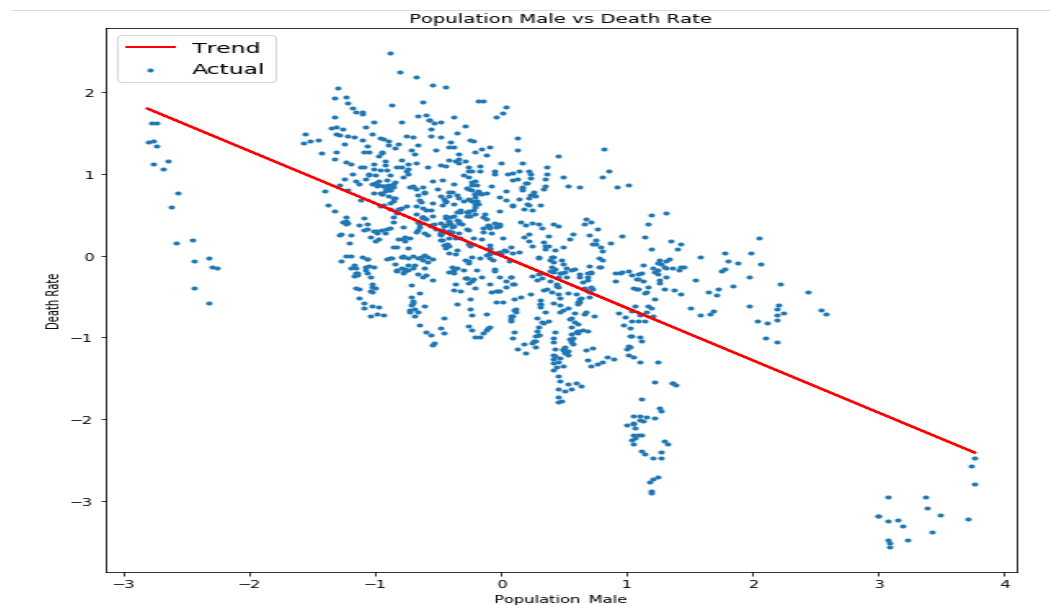
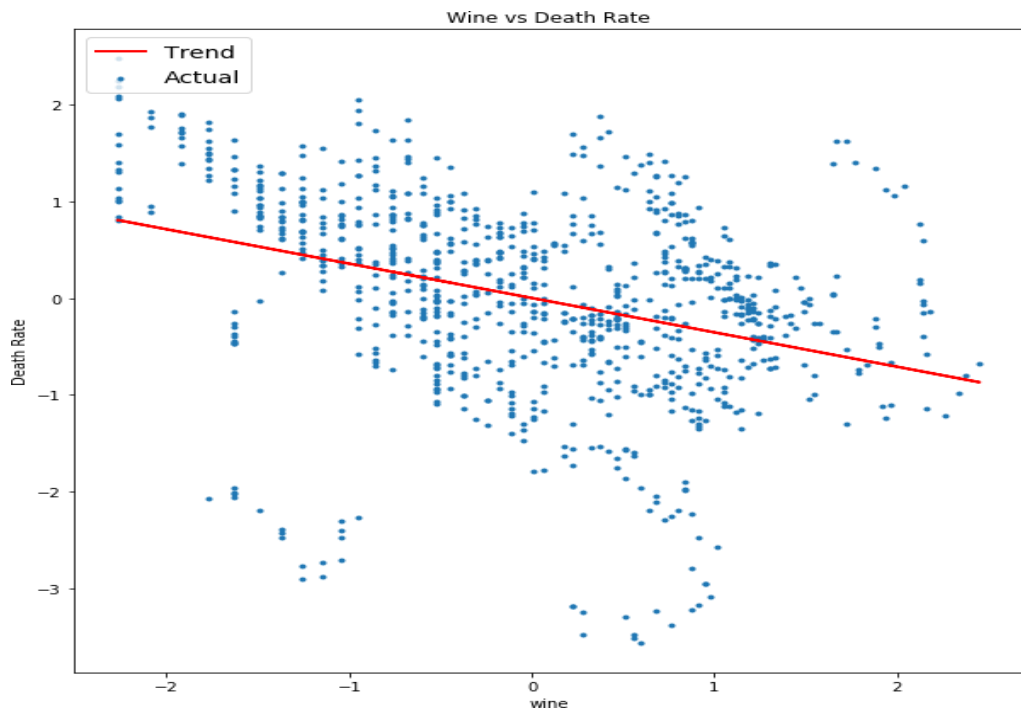
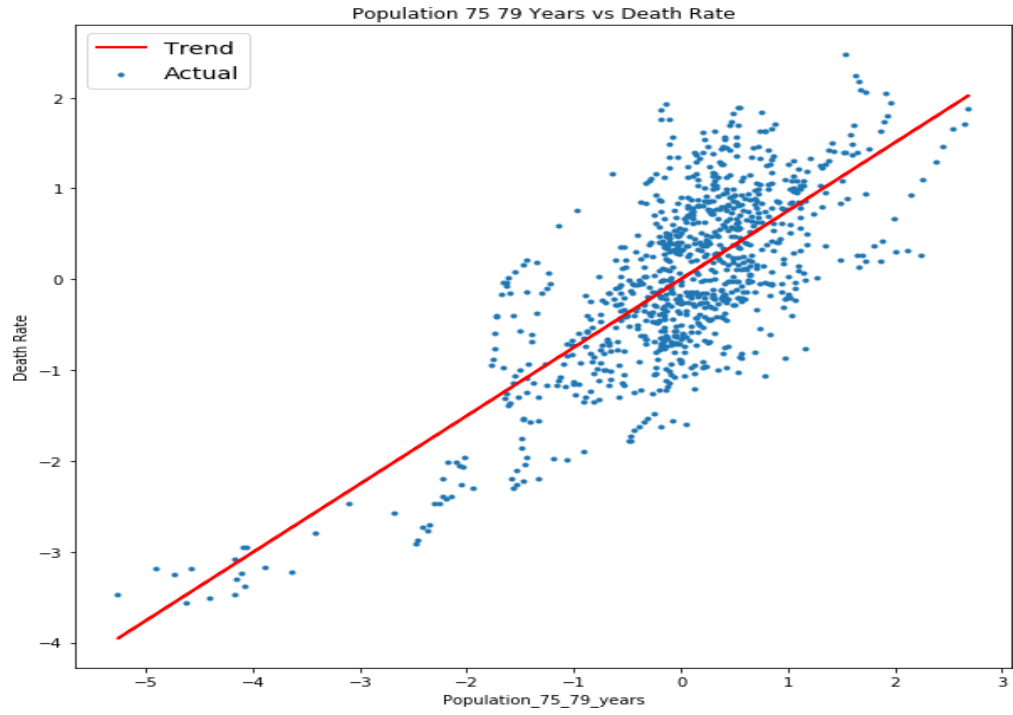
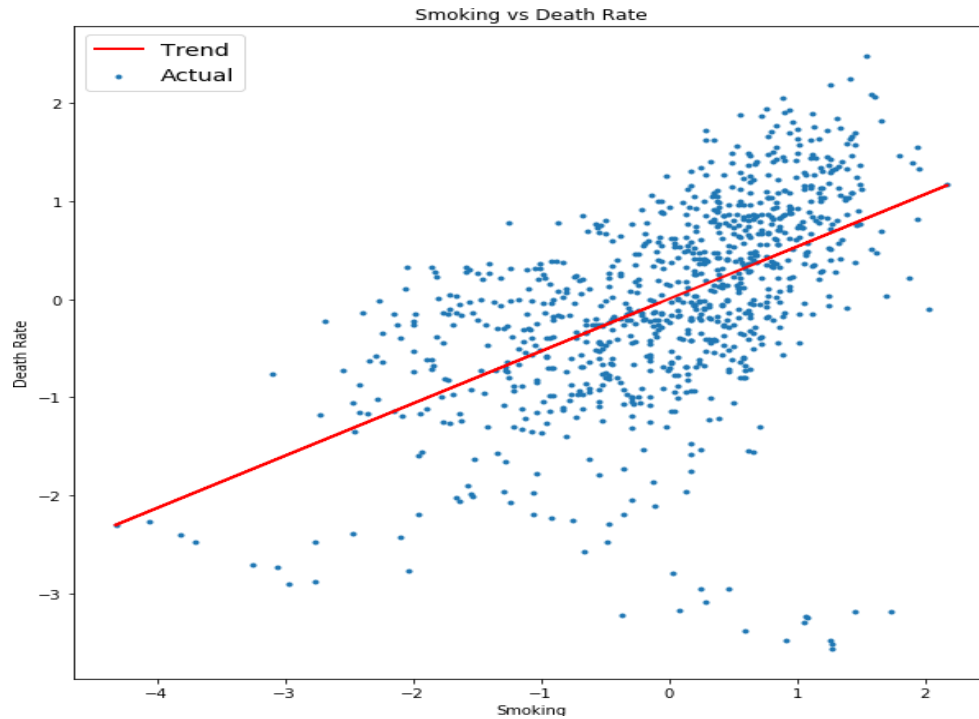


Fig 5.14 DEATHS CAUSED HEART DISEASE BY DIABETES

5.3 REGRESSION ANALYSIS







Out[62]:

	Variable	R Squared	Intercept	Slope	P Value
0	Population_Male	0.409124	2.862294e-17	-0.639628	9.108390e-107
1	Population_75_79_years	0.565296	2.862294e-17	0.751862	6.873118e-168
2	wine	0.125982	2.862294e-17	-0.354940	1.212843e-28
3	Smoking	0.283746	2.862294e-17	0.532678	2.060946e-68

Demographic data clearly has the largest effect on deaths to heart disease, particularly with regard to ages 70+ and gender. Smoking also has a clear impact, though it was surprising to see that alcohol consumption seems to have a negative relationship.

Out[70]:

	Population_Male	Population_<_1_year	Population_1_4_years	Population_5_9_years	Population_10_14_years	Population_15_19_years	Pop
State_Year							
Alabama_2016	-1.051999	-0.624540	-0.629633	-0.498151	-0.746468	-0.853802	-0.2
Alabama_2015	-1.011486	-0.529570	-0.620650	-0.429347	-0.631186	-0.878474	0.0
Alabama_2014	-0.995961	-0.612363	-0.584884	-0.374447	-0.474629	-0.901903	0.3
Alabama_2013	-0.971242	-0.728990	-0.483217	-0.328599	-0.357680	-0.771869	0.4
Alabama_2012	-0.961257	-0.545397	-0.318997	-0.339315	-0.237271	-0.626588	0.3
Alabama_2011	-0.949256	-0.431463	-0.174352	-0.342204	-0.149929	-0.343002	0.2
Alabama_2010	-0.913513	-0.391786	-0.097636	-0.229185	-0.166654	0.110520	-0.0
Alabama_2009	-0.910658	-0.004613	-0.095946	-0.150764	-0.142015	0.399590	-0.1
Alabama_2008	-0.979136	0.327165	-0.117594	-0.105159	-0.045621	0.468782	-0.1
Alabama_2007	-0.989620	0.288880	-0.169271	-0.030028	0.078014	0.432153	-0.1
Alabama_2006	-1.011595	0.090635	-0.120323	-0.015448	0.216397	0.391481	-0.0
Alabama_2005	-1.051665	0.005618	0.024407	-0.057443	0.379437	0.316569	0.1
Alabama_2004	-1.096186	0.007161	0.105738	-0.001370	0.562064	0.205659	0.1

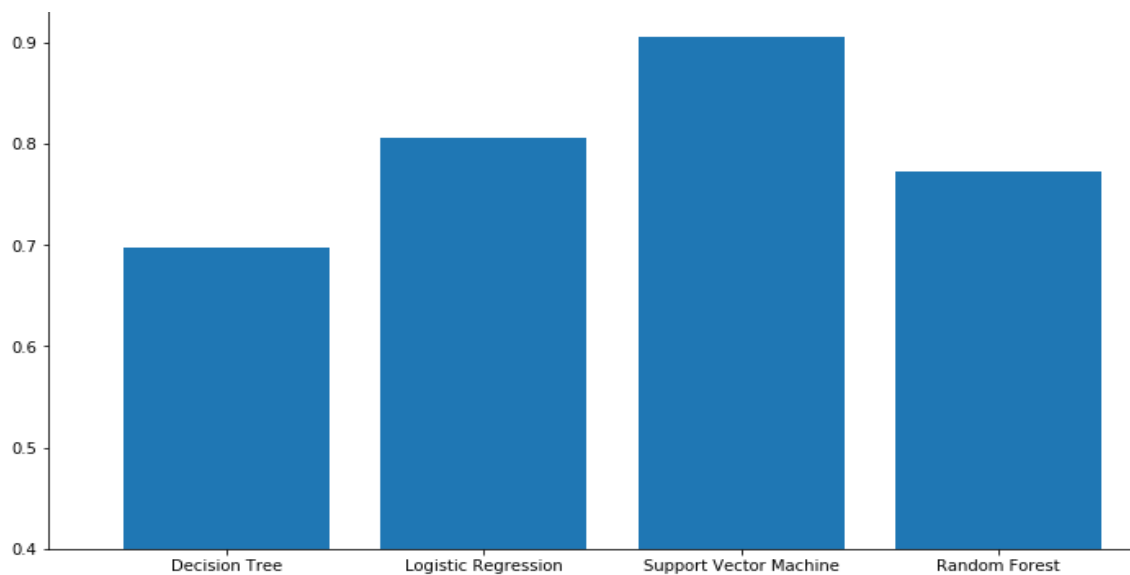
13 rows x 165 columns

A primary component analysis will first be performed to reduce dimensionality, while still maintaining 99% of variance. A 20% of the data will be withheld from the model for testing purposes, and models will be trained on the remaining data using 5 crossfold validations. Each model will be optimized to maximize the **Accuracy Score**. The follow classification models will be utilized in conjunction with Gridsearch:

- **Decision Tree Classification**
- **Logistic Regression**
- **Support Vector Machine**
- **Random Forest Classification**

Out[93]:

	Pipeline	Top Score	Parameters
0	Decision Tree	0.698057	{'clf__criterion': 'gini', 'clf__max_depth': 9...
1	Logistic Regression	0.806424	{'logreg__C': 100, 'logreg__penalty': 'l1'}
2	Support Vector Machine	0.906083	{'svm__C': 10, 'svm__degree': 1, 'svm__gamma':...
3	Random Forest	0.773339	{'rf__criterion': 'entropy', 'rf__max_depth': ...



The **Support Vector Machine Model** was much more efficient, performed better in cross validations, and performs with a high level of accuracy.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

Modelling data using machine learning techniques is done in this project. The process of using a machine to predict future events based on known past events. Firstly, from the prediction we can conclude that top causes of death are HEART DISEASE AND CANCER ,but mostly by heart disease. Then we used different machine learning models for predicting the best method to find cause of death . Based on the accuracies which we got from the machine learning algorithms we get to know that Support Vector machine is best method as well as a little bit ahead of the logistic regression method .

We expect that this work will be more useful and helpful for the future events such as taking precautions for the future needs and also helpful for the insurance companies which can set particular cost for different causes of death.

6.2 FUTURE WORK

In the future we can use different big data techniques and other programming techniques to find the cause of death which can give accurate and fast results .Those can be used for the government for taking necessary precautions

REFERENCES

- [1] “Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study” Author links open overlay panel ProfChristopher JLMurrayDPhil^aAlan DLopezPhD^b
- [2] “Analysis of Data From the World Health Organization and Coronary Artery Disease Risk Factors From NCD Risk Factor Collaboration” Alexandra N. Nowbar, Mauro Gitto, James P. Howard, Darrel P. Francis, Rasha Al-Lamee Originally published 4 Jun 2019
- [3] “Cancer statistics, 2019 “ Rebecca L. Siegel MPH ,Kimberly D. Miller MPH ,Ahmedin Jemal DVM, PhD First published: 08 January 2019
- [4] “Counting the dead and what they died from: an assessment of the global status of cause of death data” Colin D. Mathers^{1, 1}; Doris Ma Fat¹; Mie Inoue¹; Chalapati Rao¹¹; Alan D. Lopez¹¹ ¹Evidence and Information for Policy, World Health Organization, 1211 Geneva 27, Switzerland
¹¹School of Population Health, University of Queensland, Brisbane, Australia
- [5] “Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review” Sasmita Poudel Adhikari, Sha Meng, Yu-Ju Wu, Yu-Ping Mao, Rui-Xue Ye, Qing-Zhi Wang, Chang Sun, Sean Sylvia, Scott Rozelle, Hein Raat & Huan Zhou *Infectious Diseases of Poverty* volume 9, Article number: 29 (2020)
- [6] “The Swedish cause of death register” Hannah Louise Brooke, Mats Talbäck, Jesper Hörnblad, Lars Age Johansson, Jonas Filip Ludvigsson, Henrik Druid, Maria Feychting & Rickard Ljung *European Journal of Epidemiology* volume 32, pages765–773(2017)
- [7] Ahmedin Jemal, DVM, PhD; ElizabethWard, PhD; Yongping Hao, PhD, “Trends in the Leading Causes of Death in the United States, 1970-2002”, September 14,2005.
- [8] Amy L. Olson , Jeffrey J. Swigris , Dennis C. Lezotte , Jill M. Norris , Carla G. Wilson , and Kevin K. Brown, “Mortality from Pulmonary Fibrosis Increased in the United States from 1992 to 2003”, *American Journal of Respiratory and clinical care* , volume 176 , Issue 3.
- [9] Eloi Marijon, Jean-Yves Le Heuzey, Stuart Connolly, Sean Yang, Janice.Pogue, Martina Brueckmann, John,Eikelboom, Ellison,Themeles, Michael.Ezekowitz,Lars Wallentin, and Salim Yusuf, “Causes of Death and Influencing Factors in Patients With Atrial Fibrillation”,*Circulation*. 2013;128:2192–2201
- [10] Hamid Reza Hassanzadeh, Student Member, IEEE, Ying Sha, and May D. Wang, Senior Member, “DeepDeath: Learning to Predict the Underlying Cause of Death with Big Data” , Accepted for publication in the 39th IEEE

EMBC Conference (2017).

- [11] Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association Emelia J. Benjamin, Paul Muntner, Alvaro Alonso, Marcio S. Bittencourt, and On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee Originally published 31 Jan 2019
- [12] Hesham Abdo Ahmed Aqlan, Shoiab Ahmed, Ajit Danti , “Death Prediction and Analysis Using Web Mining Techniques” , 2017 International Conference on Advanced Computing and Communication Systems (ICACCS - 2015), Jan. 06 – 07, 2017, Coimbatore, INDIA.
- [13] Jing Fang, MD; Michael H. Alderman, MD,” Serum Uric Acid and Cardiovascular Mortality The NHANES I Epidemiologic ,Follow-up study 1971-1992, *JAMA*. 2000;283(18):2404-2410. doi:10.1001/jama.283.18.2404
- [14] L. Grigoletti¹ *, G. Perini¹ , A. Rossi¹ , A. Biggeri² , C. Barbui¹ , M. Tansella¹ and F. Amaddeo¹, “Mortality and cause of death among psychiatric patients: a 20-year case-register study in an area with a community-based system of care” , *Psychological Medicine* (2009), 39, 1875–1884
- [15] Laura E Caulfield, Mercedes de Onis, Monika Blössner, Robert E Black . “Undernutrition as an underlying cause of child deaths associated with diarrhea, pneumonia, malaria, and measles”, *The American Journal of Clinical Nutrition*, Volume 80, Issue 1, July 2004, Pages 193–198

APPENDIX

A.SOURCE CODE :

```
import time
T0 = time.time()

import warnings
warnings.simplefilter('ignore')

import pandas as pd
import io
import requests

## Request data via http get, then save as CSV
#url="https://data.cdc.gov/resource/bi63-dtpu.csv"
"s=requests.get(url).content
#pd.read_csv(io.StringIO(s.decode('utf-8'))).to_csv('data_files/causes_of_death.csv')

# Create dataframe from CSV
raw_df =
pd.read_csv('data_files/causes_of_death.csv')
raw_df.head()

raw_df.shape

df = raw_df[(raw_df['Cause Name'] != 'All causes') & (raw_df['State'] !=
'United States')]
df = df.drop(columns = ['113 Cause Name', 'Age-adjusted Death Rate'])
df = df.rename(columns={'Cause Name': 'Cause_Name'})
df.head()

population = pd.read_csv('data_files/Bridged-Race Population Estimates
1999-2017.csv', nrows=36822)
```

```
population.head()
```

```
"""#### Preprocessing
```

```
Initial cleanup before joining with with main dataset.
```

```
"""
```

```
population = population.drop(columns=['Notes', 'Yearly July 1st Estimates  
Code', 'State Code', 'Age Group Code', 'Gender'])
```

```
population = population.rename(columns = {'Yearly July 1st Estimates':  
'Year', 'Gender Code': 'Male'})
```

```
population['Male'] = population['Male'].replace(to_replace = {'F':0,  
'M':1}).astype(bool)
```

```
population['Year'] = population['Year'].astype(int)
```

```
population.head()
```

```
tot_pop = population.groupby(['Year',  
'State']).sum().reset_index().drop(columns=['Male'])
```

```
tot_pop.head()
```

```
add_pop_df = pd.merge(df, tot_pop, left_on=['Year', 'State'],  
right_on=['Year', 'State'])
```

```
add_pop_df.head()
```

```
# Add male population
```

```
mf_df = population[population['Male'] == True]
```

```
mf_df = mf_df.groupby(['Year',  
'State']).sum().reset_index().drop(columns=['Male'])
```

```
add_mf_df = pd.merge(add_pop_df, mf_df,
```

```
left_on=['Year', 'State'],
```

```
right_on=['Year', 'State'],
```

```
suffixes=('_Total', '_Male'))
```

```
add_mf_df['Population_Male'] = add_mf_df['Population_Male'] /
```

```
add_mf_df['Population_Total']
```

```
add_mf_df[add_mf_df['Cause_Name']=='Cancer'].head()
```

```
# Add age groups
```

```
add_age_df = add_mf_df
```

```

for group in population['Age Group'].unique():
    age_df = population[population['Age Group'] == group]
    age_df = age_df.groupby(['Year',
                             'State']).sum().reset_index().drop(columns=['Male'])
    add_age_df = pd.merge(add_age_df, age_df,
                           left_on=['Year', 'State'],
                           right_on=['Year', 'State'])
    add_age_df['Population'] = add_age_df['Population'] /
    add_age_df['Population_Total']
    add_age_df =
    add_age_df.rename(columns={'Population':f'Population_{group}'})

add_age_df[add_age_df['Year']==2016].head()

# Update column names so properly formatted for OLS functions
to_replace = {}
for col in add_age_df.columns:
    to_replace[col] = col.replace('-', '_').replace(' ', '_')

to_replace['Population_Total']='Population'
add_age_df = add_age_df.rename(columns=to_replace)
add_age_df.head()

df = add_age_df
df.info()

df.describe()

top_causes = df[['Deaths', 'Population']].groupby(df['Cause_Name']).sum()
top_causes['Deaths_per_Capita'] = top_causes['Deaths'] /
top_causes['Population']
top_causes

import matplotlib.pyplot as plt

```

```
plt.figure(figsize=(20, 7))
```

```
plt.bar(top_causes.index, top_causes['Deaths_per_Capita'])
plt.show()
```

```
def dist_plot(control, test, ctrl_label='Control Group', test_label='Test
Group'):
```

```
    # Create a distribution plot to compare control and test groups for
hypothesis testing
```

```
    plt.figure(figsize=(8, 8))
    sns.distplot(control, color='red', label=ctrl_label)
    sns.distplot(test, color='blue', label=test_label)
    plt.legend()
    plt.show()
```

```
from __future__ import print_function, division
import scipy.stats
```

```
def evaluate_PDF(rv, x=4):
```

```
    """Input: a random variable object, standard
    deviation output : x and y values for the normal
    distribution """
```

```
    # Identify the mean and standard deviation of random variable
    mean = rv.mean()
    std = rv.std()
```

```
    # Use numpy to calculate evenly spaced numbers over the specified
    # interval (4 sd) and generate 100 samples.
    xs = np.linspace(mean - x*std, mean + x*std, 100)
```

```
    # Calculate the peak of normal distribution i.e. probability density.
    ys = rv.pdf(xs)
```

```
    return xs, ys # Return calculated values
```

```
def norm_dist(control, test, ctrl_label='Control Group', test_label='Test
Group'):
```

Evaluate the PDF's of control and test groups and generate

```

# normal distributions to visualize the mean and variance differences
control_mean, control_std = np.mean(control), np.std(control)
control_quant = scipy.stats.norm(control_mean, control_std)

test_mean, test_std = np.mean(test), np.std(test)
test_quant = scipy.stats.norm(test_mean, test_std)

plt.figure(figsize=(8, 8))

xs, ys = evaluate_PDF(control_quant)
plt.plot(xs, ys, label=ctrl_label, linewidth=4, color='red')
plt.axvline(x=control_mean, color='red', ls='-.', label=f'{ctrl_label} Mean:
{np.round(control_mean, 4)}')

xs, ys = evaluate_PDF(test_quant)
plt.plot(xs, ys, label=test_label, linewidth=4, color='blue')
plt.axvline(x=test_mean, color='blue', ls='-.', label=f'{test_label} Mean:
{np.round(test_mean, 4)}')

plt.xlabel('Death Rate')
plt.legend()
plt.show()

from scipy import stats

def test_results(control, test):
    return stats.ttest_ind(test, control)

def get_results(control, test, ctrl_label='Control Label', test_label='Test
Label', conf_lvl = .95):
    # Determine if the mean differences between control and test groups
    are statistically significant
    # If so, display results & distribution plots, evaluate the PDF's, and
    provide normal distribution visualizations

    ctrl_mean, ctrl_std = np.mean(control), np.std(control)
    test_mean, test_std = np.mean(test), np.std(test)

```

```

mean_diff = (test_mean - ctrl_mean) / ctrl_mean
std_diff = (test_std - ctrl_std) / ctrl_std

t_stat, p_val = test_results(control, test)

# Print all key metrics for the test group if results are
statistically significant, along with relevant plots.
# Otherwise, indicate that the results are not significant at the
given confidence level.
print(f'Key Metrics for {test_label}')
if p_val < (1-conf_lvl):
    print(f' {ctrl_label} Mean: {np.round(ctrl_mean, 6)}')
    print(f' {test_label} Mean: {np.round(test_mean, 6)}')
    print(f' Percent Mean Difference: {np.round(100 * mean_diff,
2)}}%\n')
    print(f' {ctrl_label} Std Dev: {np.round(ctrl_std, 6)}')
    print(f' {test_label} Std Dev: {np.round(test_std, 6)}')
    print(f' Percent Std Dev Difference: {np.round(100 * std_diff, 2)}}%')
    print(f' Test Statistic: {np.round(t_stat, 4)}')
    print(f' P-Value: {p_val}')
    dist_plot(control, test, ctrl_label=ctrl_label, test_label=test_label)
    norm_dist(control, test, ctrl_label=ctrl_label, test_label=test_label)
else:
    print(f'{test_label} Not Significant at {np.round(100*conf_lvl, 2)}}%
Confidence Level\n')

# Regardless of significance, return the key statistics to a dictionary
for evaluation.
keys = ['Ctrl Mean', 'Ctrl Std', 'Test Mean', 'Test Std', 'Mean Diff', 'Std
Diff', 't-stat', 'p_val']
values = [ctrl_mean, ctrl_std, test_mean, test_std, mean_diff, std_diff,
t_stat, p_val]
return dict(zip(keys, values))
causes = df['Cause_Name'].value_counts().index
cancer = df[df['Cause_Name'] == 'Cancer']['Deaths',

```

```

'Population']].groupby(df['Year']).sum()
cancer['Deaths_per_Capita'] = cancer['Deaths'] / cancer['Population']
cancer = cancer.drop(columns = ['Deaths', 'Population'])

hd = df[df['Cause_Name'] == 'Heart disease'][['Deaths',
'Population']].groupby(df['Year']).sum()
hd['Deaths_per_Capita'] = hd['Deaths'] / hd['Population']
hd = hd.drop(columns = ['Deaths', 'Population'])

hd

cancer.head()

import numpy as np
import seaborn as sns

cancer_hd_results = get_results(cancer['Deaths_per_Capita'],
hd['Deaths_per_Capita'], ctrl_label = 'Cancer', test_label='Heart Disease')

fig = plt.figure(figsize=(15, 8))
ax = fig.gca()

xs = hd.index
ys = hd['Deaths_per_Capita']

ax.bar(xs, ys, alpha=.7)
ax.plot(xs, ys, color='r')

plt.xticks(list(range(1999, 2017, 1)))
plt.xlabel('Year')
plt.ylabel('Deaths per Capita')
plt.title('National Deaths per Capita Over Time')

plt.show()

```

```
hd_yoy = df[df['Cause_Name'] == 'Heart disease'][['Year', 'State', 'Deaths',  
'Population']]
```

```
hd_yoy['Deaths_per_Capita'] = hd_yoy['Deaths'] / hd_yoy['Population']
```

```
hd1999 = hd_yoy[hd_yoy['Year'] == 1999]['Deaths_per_Capita']
```

```
hd2016 = hd_yoy[hd_yoy['Year'] == 2016]['Deaths_per_Capita']
```

```
hd_results = get_results(hd1999, hd2016, ctrl_label = 'Heart Disease  
1999', test_label='Heart Disease 2016', conf_lv=.95)
```

```
"""<a id='methodology'></a>
```

```
## Model Methodology
```

```
* **Decision Tree Classification**
```

```
* **Logistic Regression**
```

```
* **Support Vector Machine**
```

```
* **Random Forest Classification**
```

```
<a id='methodology'></a>
```

```
### Identify Top Features with PCA
```

```
"""
```

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(features, target,  
test_size=.2, random_state=123)
```

```
from sklearn.decomposition import PCA
```

```
pca = PCA(n_components=.99, random_state=123)
```

```
feat_pca = pca.fit_transform(features)
```

```
n_feat = feat_pca.shape[1]
```

```
print(f'Number of Features: {n_feat}\n')
```

```
print(f'Cumulative Sum
```

```
Variance:\n{np.cumsum(pca.explained_variance_ratio_)})
```

```
"""<a id='tree'></a>
```

```
### Decision Tree Pipeline with PCA & Gridsearch
```

|||||

```

from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report

tree_pipe = Pipeline([('pca', PCA(n_components=n_feat,
                                random_state=123)),
                      ('clf', DecisionTreeClassifier(random_state=123))])

tree_grid = [{'clf__criterion': ['gini', 'entropy'],
              'clf__max_depth': list(range(1, 11, 2)),
              'clf__min_samples_split': [i/100 for i in range(5, 96, 10)],
              'clf__min_samples_leaf': list(range(1, 11, 2))}]

acc_meas = 'balanced_accuracy'
tree_gridsearch = GridSearchCV(estimator = tree_pipe,
                               param_grid = tree_grid,
                               scoring = acc_meas,
                               cv=5, verbose = 1, return_train_score = True)

tree_gridsearch.fit(X_train, y_train)
print(f'Best Decision Tree Pipeline Score: {tree_gridsearch.best_score_}')
print(f'Parameters: {tree_gridsearch.best_params_}')

results = {'Pipeline': ['Decision Tree'],
          'Top Score': [tree_gridsearch.best_score_],
          'Parameters': [tree_gridsearch.best_params_]}

"""<a id='logreg'></a>
### Logistic Regression Pipeline with PCA & Gridsearch
"""

from sklearn.linear_model import LogisticRegression

logreg_pipe = Pipeline([('pca', PCA(n_components=n_feat,

```

```
random_state=123)),  
      ('logreg', LogisticRegression(random_state=123)))]
```

```
logreg_grid = [{'logreg__C': [.001, 0.01, 0.1, 1, 10, 100, 1000],  
                'logreg__penalty': ['l1', 'l2']}
```

```
logreg_gridsearch = GridSearchCV(estimator = logreg_pipe,  
                                  param_grid = logreg_grid,  
                                  scoring = acc_meas,  
                                  cv = 5, verbose = 1, return_train_score = True)
```

```
logreg_gridsearch.fit(X_train, y_train)
```

```
print(f'Best Logistic Regression Pipeline Score:  
{logreg_gridsearch.best_score_}')  
print(f'Parameters: {logreg_gridsearch.best_params_}')
```

```
results['Pipeline'].append('Logistic Regression')  
results['Top Score'].append(logreg_gridsearch.best_score_)  
results['Parameters'].append(logreg_gridsearch.best_params_)
```

```
"""<a id='svm'></a>
```

```
### Support Vector Machine Pipeline with PCA & Gridsearch  
"""
```

```
from sklearn.svm import SVC
```

```
svm_pipe = Pipeline([('pca', PCA(n_components=n_feat,  
                                random_state=123)),  
                     ('svm', SVC(random_state=123))])
```

```
svm_grid = [{'svm__kernel': ['linear'],  
             'svm__C': [.1, 1, 10],  
             'svm_degree': list(range(1, 11, 2))},
```

```
             {'svm__kernel': ['rbf'],  
              'svm__C': [.1, 1, 10],  
              'svm_degree': list(range(1, 11, 2))},
```

```
'svm_gamma': [0.001, .01, 'auto']  
]
```

```

svm_gridsearch = GridSearchCV(estimator = svm_pipe,
                               param_grid = svm_grid,
                               scoring = acc_meas,
                               cv = 5, verbose = 1, return_train_score = True)

```

```

svm_gridsearch.fit(X_train, y_train)

```

```

print(f'Best SVM Pipeline Score: {svm_gridsearch.best_score_}')
print(f'Parameters: {svm_gridsearch.best_params_}')

```

```

results['Pipeline'].append('Support Vector Machine')
results['Top Score'].append(svm_gridsearch.best_score_)
results['Parameters'].append(svm_gridsearch.best_params_)

```

```

"""<a id='rf'></a>

```

```

#### Random Forest with PCA & Gridsearch

```

```

"""

```

```

from sklearn.ensemble import RandomForestClassifier

```

```

rf_pipe = Pipeline([('pca', PCA(n_components=n_feat,
                                random_state=123)),
                    ('rf', RandomForestClassifier(random_state=123))])

```

```

rf_grid = [{'rf__criterion': ['gini', 'entropy'],
            'rf_max_depth': list(range(1, 11, 2)),
            'rf_min_samples_split': [i/100 for i in range(5, 96, 10)],
            'rf_min_samples_leaf': list(range(1, 11, 2))}]

```

```

rf_gridsearch = GridSearchCV(estimator = rf_pipe,
                              param_grid = rf_grid,
                              scoring = acc_meas,
                              cv = 5, verbose = 1, return_train_score = True)

```

```
rf_gridsearch.fit(X_train, y_train)
```

```
print(f'Best Random Forest Pipeline Score: {rf_gridsearch.best_score_}')
print(f'Parameters: {rf_gridsearch.best_params_}')
```

```
results['Pipeline'].append('Random Forest') results['Top
Score'].append(rf_gridsearch.best_score_)
results['Parameters'].append(rf_gridsearch.best_params_)
```

```
"""<a id='modresults'></a>
```

```
## Model Results
```

```
"""
```

```
results_df = pd.DataFrame.from_dict(results)
results_df.to_csv('data_files/model_results.csv')
results_df.to_excel('data_files/model_results.xlsx')
results_df
```

```
plt.figure(figsize=(18, 7))
plt.bar(results_df['Pipeline'], results_df['Top Score'])
plt.ylim(.4, .95)
plt.show()
```

```
from sklearn.metrics import accuracy_score
```

```
# For each model, print the accuracy score and classification report.
```

```
tree_preds = tree_gridsearch.predict(X_test)
tree_accuracy = accuracy_score(y_test, tree_preds)
```

```
logreg_preds = logreg_gridsearch.predict(X_test)
logreg_accuracy = accuracy_score(y_test, logreg_preds)
```

```
svm_preds = svm_gridsearch.predict(X_test)
svm_accuracy = accuracy_score(y_test, svm_preds)
```

```
rf_preds = rf_gridsearch.predict(X_test)
rf_accuracy = accuracy_score(y_test, rf_preds)
```

```

#ada_preds = ada_gridsearch.predict(X_test)
#ada_accuracy = accuracy_score(y_test, rf_preds)

#xgb_preds = xgb_gridsearch.predict(X_test)
#xgb_accuracy = accuracy_score(y_test, xgb_preds)

# Generate model that calculates the weighted accuracy score of
# predictions. Weightings based on test set results.
tot_accuracies = tree_accuracy + logreg_accuracy + svm_accuracy +
rf_accuracy

overall_preds = np.round((np.array(tree_preds) * tree_accuracy +
                           np.array(logreg_preds) * logreg_accuracy +
                           np.array(svm_preds) * svm_accuracy +
                           np.array(rf_preds) * rf_accuracy / tot_accuracies, 0).astype(int)

overall_accuracy = accuracy_score(y_test, overall_preds)

print(f'Decision Tree Accuracy: {tree_accuracy*100:.2f}%')
print(f'-----Classification Report      \n{classification_report(y_test,
tree_preds)}\n')

print(f'Logistic Regression Accuracy: {logreg_accuracy*100:.2f}%')
print(f'-----Classification Report      \n{classification_report(y_test,
logreg_preds)}\n')

print(f'Support Vector Machine Accuracy: {svm_accuracy*100:.2f}%')
print(f'-----Classification Report      \n{classification_report(y_test,
svm_preds)}\n')

print(f'Random Forest Accuracy: {rf_accuracy*100:.2f}%')
print(f'-----Classification      Report      \n{classification_report(y_test,
rf_preds)}\n')

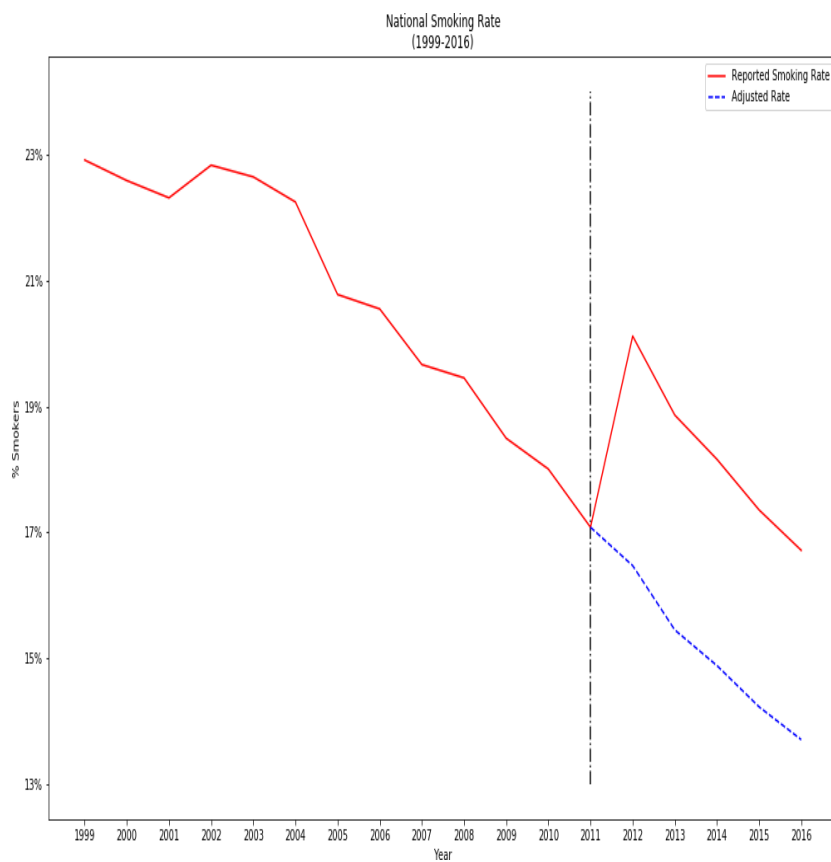
print(f'Combined Accuracy: {overall_accuracy*100:.2f}%')
print(f'-----ClassificationReport      \n{classification_report(y_test,
overall_preds)}\n')

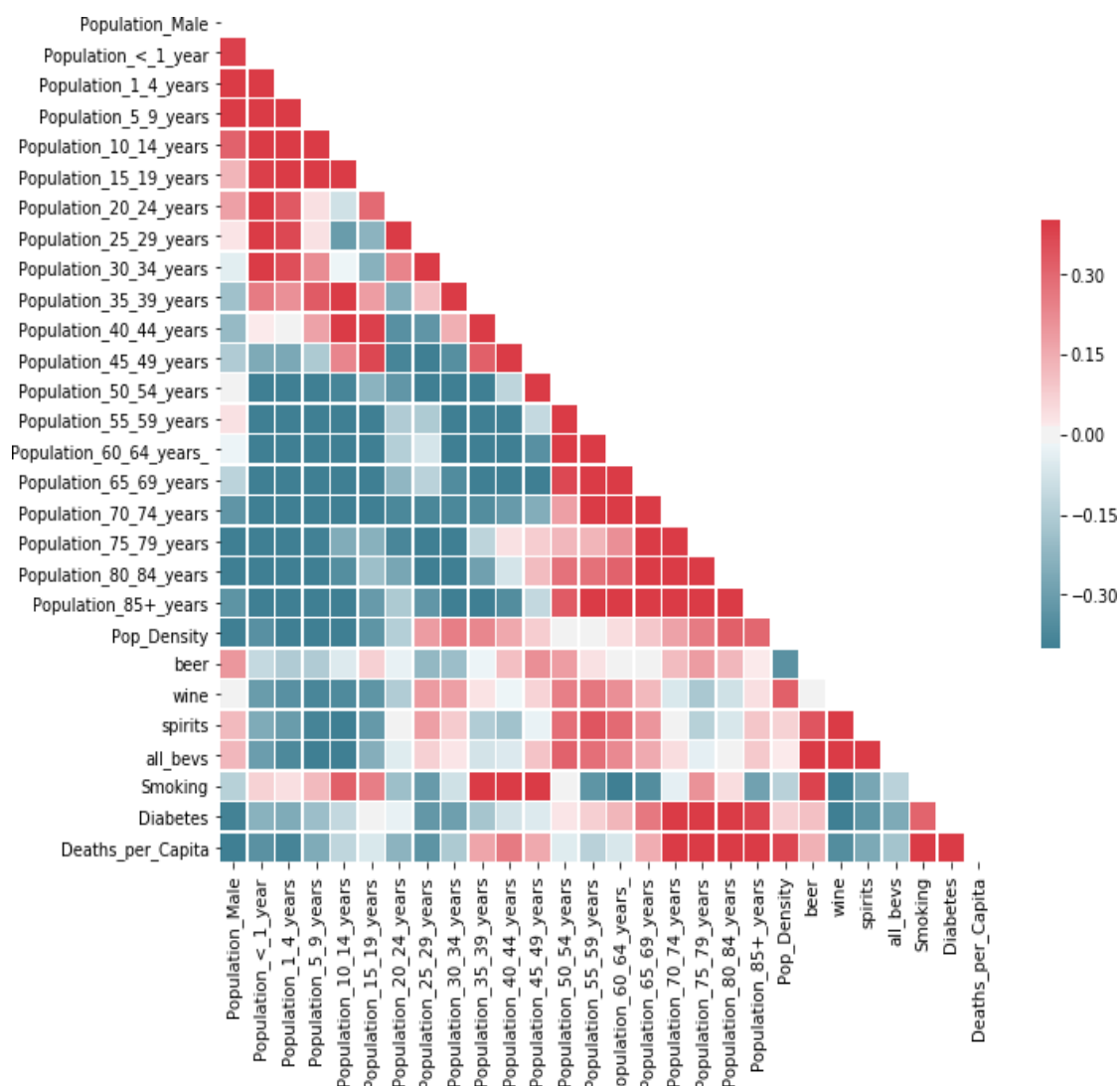
```

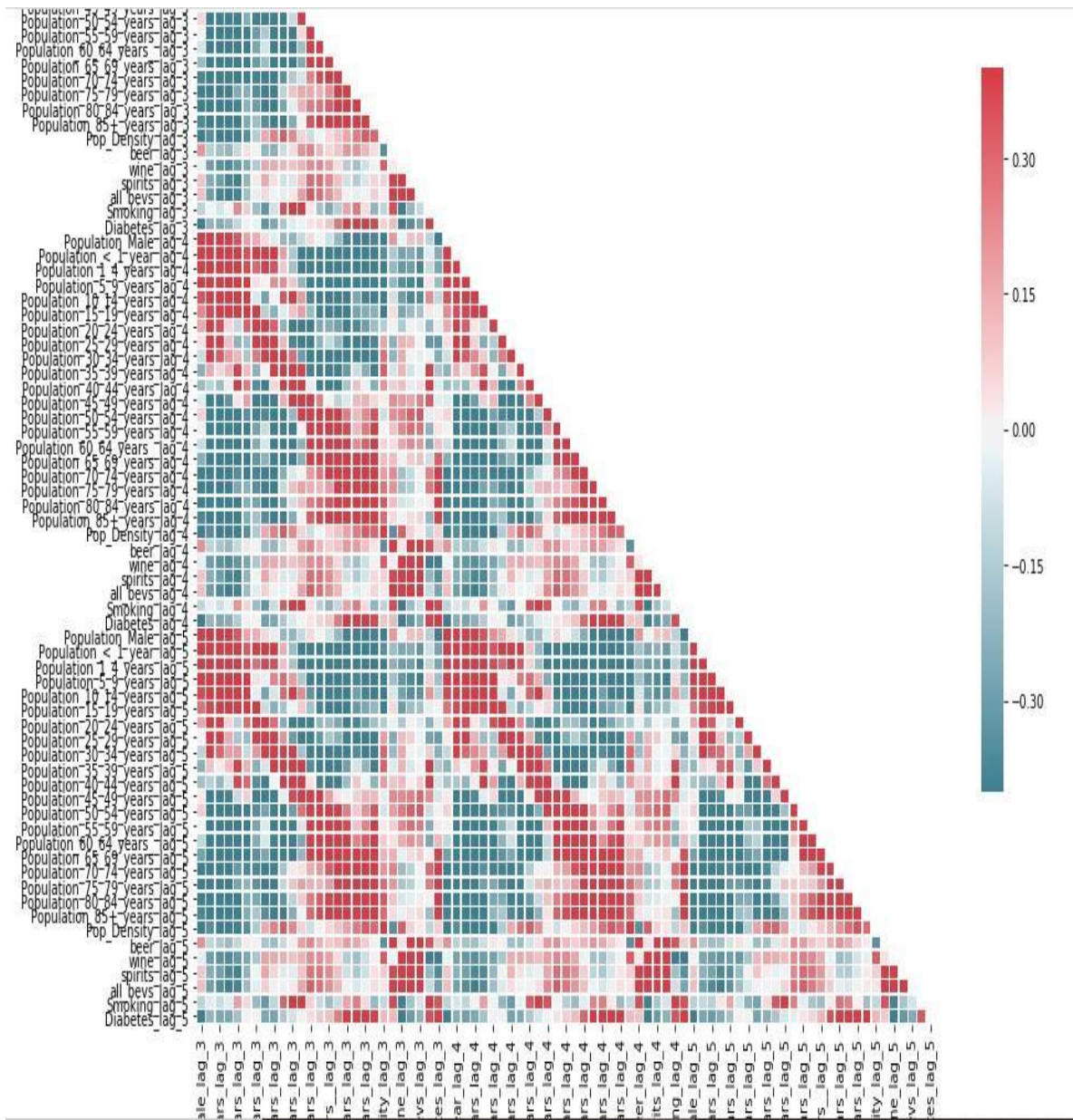
B. SCREEN SHOTS

Out[45]:

	Smoking	Smoking_unadj
Year		
2011	0.170815	0.170815
2012	0.164729	0.201165
2013	0.154513	0.188618
2014	0.148854	0.181601
2015	0.142306	0.173562
2016	0.137095	0.167155







Decision Tree Accuracy: 74.44%

-----Classification Report-----

	precision	recall	f1-score	support
1	0.82	0.87	0.84	52
2	0.67	0.64	0.65	47
3	0.73	0.71	0.72	34
micro avg	0.74	0.74	0.74	133
macro avg	0.74	0.74	0.74	133
weighted avg	0.74	0.74	0.74	133

Logistic Regression Accuracy: 76.69%

-----Classification Report-----

	precision	recall	f1-score	support
1	0.85	0.90	0.88	52
2	0.67	0.66	0.67	47
3	0.75	0.71	0.73	34
micro avg	0.77	0.77	0.77	133
macro avg	0.76	0.76	0.76	133
weighted avg	0.76	0.77	0.76	133

Support Vector Machine Accuracy: 91.73%

-----Classification Report-----

	precision	recall	f1-score	support
1	0.96	0.94	0.95	52
2	0.86	0.91	0.89	47
3	0.94	0.88	0.91	34
micro avg	0.92	0.92	0.92	133
macro avg	0.92	0.91	0.92	133
weighted avg	0.92	0.92	0.92	133

Random Forest Accuracy: 80.45%

-----Classification Report-----

	precision	recall	f1-score	support
1	0.92	0.85	0.88	52
2	0.71	0.79	0.75	47
3	0.79	0.76	0.78	34
micro avg	0.80	0.80	0.80	133
macro avg	0.81	0.80	0.80	133
weighted avg	0.81	0.80	0.81	133

CAUSE OF DEATH ANALYSIS

P.Harshika¹,G.Sowmya² and Dr.Usha Nandini³

¹B.E, Computer Science and Engineering,Sathyabama Institute of Science and Technology ,

²B.E, Computer Science and Engineering ,Sathyabama Institute of Science and Technology ,

³Faculty of computing,Sathyabama Institute of Science and Technology

ABSTRACT

Death records can offer insights into many aspects of current existence. This venture is ready to locate the predominant explanation for deaths from the dataset. The fundamental cause of this project is to make consciousness approximately the deaths in everywhere in the country . So that their authorities can take important precautions for destiny and is likewise useful for coverage organization to set positive quantity for specific dying. The machine used for this project is getting to know algorithms and python programming. This paper proposes approximately predicting the reasons for dying in a specific period through thinking about the entities withinside the statistics set. The end result offers the accuracy for all of the anticipated system getting to know fashions.

KEYWORDS- Death Statistics, Machine Learning, python programming.

INTRODUCTION

Machine Learning Algorithms are primarily used for predicting and analyzing from a huge dataset very easily. Predicting

the destiny through units of statistics from the admitted sufferer's documents in a sanatorium from the years 1999-2016 is defined and predicting occasions from the existing dataset. Different system machine learning techniques to get to know strategies might be used for predicting. The cause of death is an reliable will power of situations ensuing in a people dying , which can be recorded on a dying certificate. The reason of dying can be of unique ailment or harm, in assessment to the way of dying that's small number of classes like "natural"," Twist of fate"," suicide", and "murder" that have unique criminal implications. Cause of dying additionally gives data public fitness ,making plans and policy. Causes of dying statistics can tell thoughts as to what preventative measures the metropolis need to put money into enhances stardiness withinside the metropolis.

Today, evaluation of mortality statistics more and more more makes use of a couple of reason of dying method. Generally, reason of dying records is derived from the so known as underlying reason of dying in a manner hereafter refed to because the traditional approach. Department of Health and Mental Hygiene (DOHMH) defines the underlying reason of dying because "the ailment or harm which initiated the occasions main at once to dying or the situations of the twist of a fate or violence which produced deadly effects. However, deaths are regularly caused as a result

of multiple ailments. Moreover, in a global characterized through an getting old population and reducing mortality and fertility. Death because of infectious ailment is regularly being changed through dying because of continual and degenerative illness.

RELATED WORK

Jing Fang et al.,[1] built Cox regression models for locating greater quantity of deaths through aerobic vascular ailments in males and female individually. Laura E Caulfield et al.,[2] innovated the cause of dying in youngsters due to underweight and additionally anticipated the fraction of deaths through attributes to underweight through the usage of the regression analysis. Ahmedin Jemal et al.,[3] observed the traits of main reason of deaths withinside the years 1970-2002 of USA through the usage of regression evaluation. And additionally, observed the mortality of coronary heart ailment in unique aspects. Amy et al.,[4] have calculated the age smart mortality changes from the deaths as a result of pulmonary fibrosis and described the statistics to decide mortality rate variations through gender, race and so on through the usage of Poisson multivariable regression approach. L. Rigolet et al.,[5] observed the quantity of mortality and additionally the reason of dying many of the psychiatric sufferers through the usage of Poisson regression strategies. Jean-Yves Le Heuzey et al.,[6] observed the reason of dying and additionally the influencing elements of the sufferers with atrial fibrillation. They have used the path layout approach and additionally the statistical evaluation for getting the correct outcomes. Hesham Abdo Ahmed Aqlan et al.,[7] have accomplished dying prediction the usage of web mining techniques. They have compared all the methods and observed the satisfactory approach with maximum accuracy as end result. Hamid Reza Hassanzadeh et al.,[8] observed the underlying causes of death through the usage of Big Data techniques. They even used n-Gram models to

understand associations among multiple cases.

EXISTING SYSTEM

It is very much important to know the cause of death for people. Knowing the cause of death would be a great help for health and life insurance companies to know as they can make decisions on what prices to set for different kinds of insurance and coverage. In olden days, to realize the reason of dying, we need to personally go to the door-to-door survey for enquiry and need to make a list of deaths which is a time-consuming process. But now we can store the information once they get admitted in any hospital which is then recorded in the form of database. From the base paper and other research papers observed we get to realize the unique approaches of finding the causes of death.

Later on, prediction of deaths became accomplished through the usage of classification algorithms like Multilayer Perceptron(MLP), Linear Regression, SMO Regression, MLP Regressor. The fundamental goal of these classification algorithms to predict the future incidents through the usage of classifier[7]. In MLP approach back propagation neural network with multilayer perceptron is used to classify instances and it could be changed during training time. In Linear Regression weighted instances are used withinside the given equation for future forecast of deaths. But those algorithms are very tedious for large datasets. Some alternative integrated methods are used for finding the reason of dying for existence expectancy. When measured for a specific period, life expectancy at birth is a precis measure of the mortality conditions of that period through the usage of decomposition approach al.,[9].

PROPOSED METHODOLOGY

In this proposed approach, statistics is extracted for prediction of death event and later preprocessed and differentiated numerous machine learning algorithms on reasons of dying dataset of USA. Then diagnosed the satisfactory overall performance set of rules for prediction and forecasting of reason for dying illness.

The statistics we used for this project is from the Center for Disease Control's (CDC) National Center for Health Statistics (NCHS). They preserve the database with age-wise death rates and counts throughout the united states. The statistics is grouped by cause, year, state from the years 1999-2016.

Data preprocessing is a manner of preparing the raw data and making it appropriate for a machine learning model. It is the first and vital step while creating a machine learning model. When developing a machine learning project, it is not continually a case that we come upon the clean and formatted statistics. And while doing any operation with statistics, it is obligatory to clean it and put in a formatted manner. So for this, we use data preprocessing task. Data preprocessing is required obligations for cleaning the data and making it appropriate for a machine learning model which additionally increases the accuracy and performance of a machine learning model.

This project includes different phases or modules. They are:

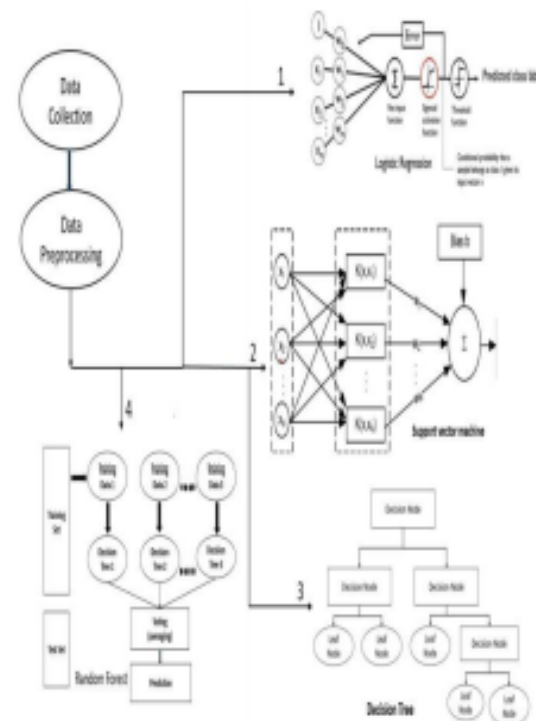
Module 1 : In this step the data will be read from the csv files which are taken from the Center for Disease Control's (CDC) National Center for Health Statistics (NCHS) and are uploaded. Then the preprocessing of the dataset is done in which data gets cleaned and even removes columns if required.

Module 2 : In this module the data is transformed and gets ready for the prediction. Then we will consider few factors and find them. They are

- 1) The main leading cause of dying in the USA
- 2) The current trend for the main leading cause of death
- 3) Which factors have the maximum impact on these deaths

Based on these factors we will find the major cause of death in USA.

Module 3 : In this module we will be using different machine learning algorithms for prediction. And based on the results of all the algorithms we will conclude the best method for predicting using their accuracy values.



Architecture Diagram

MODULES DESCRIPTION

In the module 1 , we will perform preprocessing of data. As preprocessing of the dataset is the most important step for any machine learning project. It includes

- Collection of datasets
- Cleaning the dataset

In the very next module, we will be finding the major leading cause of death in United States. Firstly , the data is grouped by top causes of death .From that result we get to know that the cause of death is "Heart disease" and a little difference of "Cancer". So, we performed Hypothesis Testing to find the difference in death rates between Heart disease and Cancer. Then we created data frames for comparing the both causes and found that heart disease is causing majorly ahead of cancer.

The very next thing we wanted to know is the major factors causing the death by heart disease. So, we considered few factors such as

- Population Density
- Age Demographic
- Gender Demographic
- Alcohol Consumption
- Smoking Population
- Diabetes Rates

For each factor the data is scraped and converted to csv file and then the process is performed .Then we performed regression analysis for all the factors which are considered for Heart disease.

Linear Regression :

Linear Regression is a supervised machine learning algorithm .It is a statistical approach that is used for predictive evaluation. Linear regression

algorithm shows a linear relationship among a dependent (y) and one or more independent (x) variables with a sloped straight line, consequently known as linear regression. It observes how the value of the dependent variable is changing according to the value of the independent variable.

The equation has the form $Y = a + bX$, where Y is the dependent variable (that is the variable which goes on the Y axis), X is the independent variable (it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$
$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Finding the appropriate method :

1) Logistic Regression : It is a form of regression that lets in the prediction of discrete variables through a combination of continuous & discrete predictors.

In logistic regression , the final results variable is binary and the purpose of the evaluation is to assess the effects of multiple explanatory variables which may be and/or categorical.

The PCA does an unsupervised dimensionality reduction, while the logistic regression does the prediction. We use a GridSearchCV to set the dimensionality of the PCA.

$$P = e^{a+bX} / 1 + e^{a+bX}$$

is the equation used for logistic regression. In the result we get the accuracy as 0.8064 2373817349.

2) Support Vector Machine : These are the supervised learning models with associated learning algorithms that analyse data and recognise patterns used for regression analysis. I have combined SVM, PCA, and Grid-search to create a pipeline to find best parameters for binary classification & eventually plot a decision boundary to present how good our algorithm has performed.

As the result it achieved a 90% accuracy with all cross-sections of test data.

3) Decision Tree : A decision tree is a decision support tool that uses a tree-like model of choice and their feasible consequences, including chance event results, resource costs, and utility. It is one manner to display an algorithm that most effectively contains conditional control statements. In this the data is spilt into few parameters.

In this information gain is used and then Data gets divided into separate rows. As the result we get the best decision tree pipeline score accuracy of 69 % .

4) Random Forest Method : In this method it selects some random samples from the data and constructs decision trees for each sample and gets a result from each decision tree .Then it performs voting and predicts the final result .It can be used for both classification and also for regression problems. If there are a greater number of trees in forest then the accuracy will be highest. It considers the highest voted target as the final prediction .

In this we got the best random forest pipeline score with accuracy of 77% .

RESULT

Modelling data using machine learning techniques . The process of using a machine to predict future events based on

known past events .The data is used for predicting the past analysis which may be helpful for the future . In data processing 17 year of data set has been considered. From this the maximum causes of death for all diseases have been computed. And also found the main cause of major deaths i.e., Heart disease. Through which we found that males of age above 70 are affected the most .

Then we proposed machine learning algorithms for finding the best way to predict the heart disease . Such as using Decision Tree algorithm, Random Forest method , Support Vector Machine and Logistic Regression. The results of all the algorithms have been evaluated and found that Support vector machine algorithms got the highest accuracy. So, the predictions says that it is the best method to find the heart disease death analysis.

93]:

	Pipeline	Top Score	Parameters
0	Decision Tree	0.658057	(crt_criterion: 'gini', 'crt_max_depth: 9...
1	Logistic Regression	0.806424	(logreg_C: 100, 'logreg_penalty: 11')
2	Support Vector Machine	0.906083	('svm_C': 10, 'svm_degree': 1, 'svm_gamma':...
3	Random Forest	0.773339	('rf_criterion: 'entropy', 'rf_max_depth: ...

CONCLUSION

Modelling data using machine learning techniques . The process of using a machine to predict future events based on known past events .Based on the accuracies which we got from the machine learning algorithms we get to know that Support Vector machine is best method as well as a little bit ahead of the logistic regression method .

We expect that this work will be more useful and helpful for the future events such

as taking precautions for the future needs and also helpful for the insurance companies which can set particular cost for different causes of death.

REFERENCES

- [1] Jing Fang, MD; Michael H. Alderman, MD ,” Serum Uric Acid and Cardiovascular Mortality The NHANES I Epidemiologic Follow-up study 1971-1992,” *JAMA*. 2000;283(18):2404-2410.
doi:10.1001/jama.283.18.2404
- [2] Laura E Caulfield, Mercedes de Onis, Monika Blössner, Robert E Black . “Undernutrition as an underlying cause of child deaths associated with diarrhea, pneumonia, malaria, and measles”, *The American Journal of Clinical Nutrition*, Volume 80, Issue 1, July 2004, Pages 193–198
- [3] Ahmedin Jemal, DVM, PhD; Elizabeth Ward, PhD; Yongping Hao, PhD, “Trends in the Leading Causes of Death in the United States, 1970-2002”, September 14, 2005.
- [4] Amy L. Olson , Jeffrey J. Swigris , Dennis C. Lezotte , Jill M. Norris , Carla G. Wilson , and Kevin K. Brown , “Mortality from Pulmonary Fibrosis Increased in the United States from 1992 to 2003”, *American Journal of Respiratory and clinical care* , volume 176 , Issue 3.
- [5] L. Grigoletti¹ *, G. Perini¹ , A. Rossi¹ , A. Biggeri² , C. Barbuil , M. Tansella¹ and F. Amadeo¹, “Mortality and cause of death among psychiatric patients: a 20-year case-register study in an area with a community-based system of care” , *Psychological Medicine* (2009), 39, 1875–1884
- [6] Eloi Marijon, Jean-Yves Le Heuzey, Stuart Connolly, Sean Yang, Janice Pogue, Martina Brueckmann, John Eikelboom, Ellison Themeles, Michael Ezekowitz, Lars Wallentin, and Salim Yusuf, “Causes of Death and Influencing Factors in Patients With Atrial Fibrillation”, *Circulation*. 2013;128:2192–2201
- [7] Hesham Abdo Ahmed Aqlan, Shoiab Ahmed, Ajit Danti , “Death Prediction and Analysis Using Web Mining Techniques” , 2017 International Conference on Advanced Computing and Communication Systems (ICACCS -2015), Jan. 06 – 07, 2017, Coimbatore, INDIA.
- [8] Hamid Reza Hassanzadeh, Student Member, IEEE, Ying Sha, and May D. Wang, Senior Member, “DeepDeath: Learning to Predict the Underlying Cause of Death with Big Data” , Accepted for publication in the 39th IEEE EMBC Conference (2017).