Exploratory Data Analysis by Employing Statistical and Machine Learning Techniques on Cardio Vascular and Pulmonary Disorders

[1] Gosu Naveen, [2] Salina Adinarayana, [3] R. Rajeswara Rao

[1]Research scholar, Department of CSE,

Jawaharlal Nehru Technological University, Vizianagaram, Andhra Pradesh, India

[2]Supervisor, Professor, Department of CSE,

Raghu Institute of Technology(A), Visakhapatnam, Andhra Pradesh, India

[3]Co-supervisor & Professor, Department of CSE

Jawaharlal Nehru Technological University, Vizianagaram, Andhra Pradesh, India

Abstract

In today's information age, data acts as a catalyst while the data analysis plays a vital role to empower any sector. Health informatics is an area which is momentous and highly progressive. The statistical information extracted from the data is highly influential in making decisions in various perspectives. Moreover, the Cardio and respiratory disorders are constantly making a huge impact. Although the governments are making their efforts to address this global issue, its prevalence is becoming burdensome. Covid had made the situation disruptive. In this paper, a recent Covid dataset is considered to be the subject for analysis. A detailed analysis has been made on the data and the results are discussed comprehensively. Although the dataset contains data pertaining to various health disorders, the focus of paper is on Cardio vascular and pulmonary data. Various analytical aspects such as Uni variate analysis, Bi variate analysis, Hypothesis testing and Machine Learning are performed. Survival status due to various disease factors also is discussed. In this Artificial Intelligence era, Machine learning and Deep Learning are imperative and authoritative. Few machine learning analysis models are applied on the data and discussed precisely. Finally, how these statistical results could be interpreted are specified.

Keywords: *Statistical analysis on healthcare data, Healthcare data analysis, Pulmonary data analysis, Cardio vascular data analysis.*

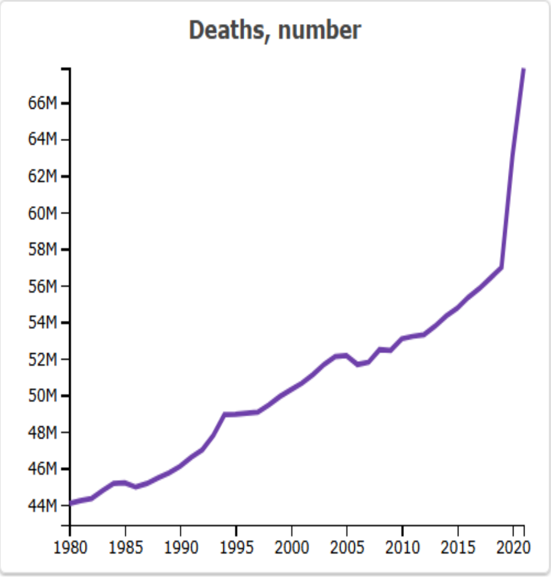
1. Introduction

Health is an intrinsic part of life and due to having good health a productive lifestyle could be led. In today's world abnormal living and food habits are leading to health disorders which are making the entire society disabled. Despite the technology advancements, the epidemics and pandemics are choking the people as well governments. Majorly the epidemic progression is characterized by the U-turn in the socio-economic gradients; usage of tobacco; consumption of low vegetables and fruits and other such factors.

Not only the biomedical approaches which are being implemented as part of treatment and diagnosis, but also the focus must be on social determinants such as Physical inactivity, usage of tobacco, harmful alcohol usage, air pollution, poverty, unhealthy diet, gender inequalities and occupational exposure and pressure so that disease prevention could be effectively done. Although several strategies are being implemented by the Governments, technology being updated; the mortality rate is exponentially increasing worldwide. The number of deaths is depicted in figure 1 below. Moreover, the top 10 causes [3] of death are also presented in figure 2.

Fig 1: No. of deaths over the time

It could be understood that among the top 10 diseases those are instigating mortality, heart diseases and pulmonary diseases are the leading death causes. Although Covid 19 seems to be a leading one, it is influential during the pandemic years. However later due to the pandemic, the vital human organs of the effected persons are at risk.



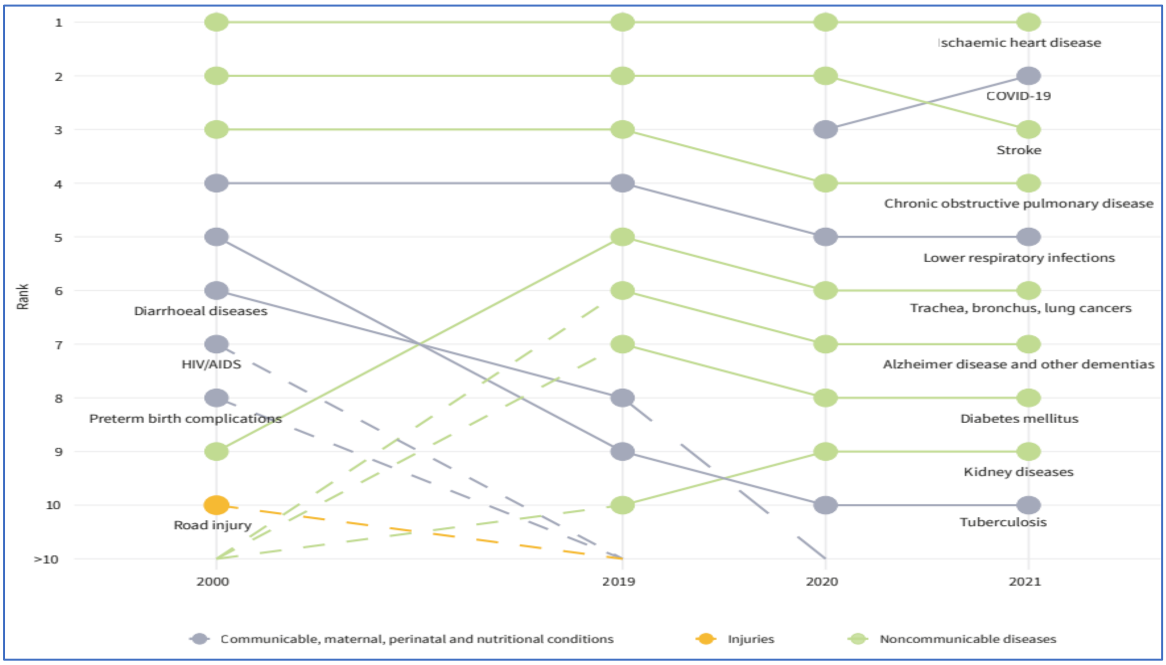


Fig 2: Top 10 causes of death (globally) – 2000, 2019, 2020, 2021.

Health diseases are basically of two categories communicable and non-communicable among which the latter ones are impacting more. The trend of communicable vs non communicable for the past twenty years is as shown in figure 3.

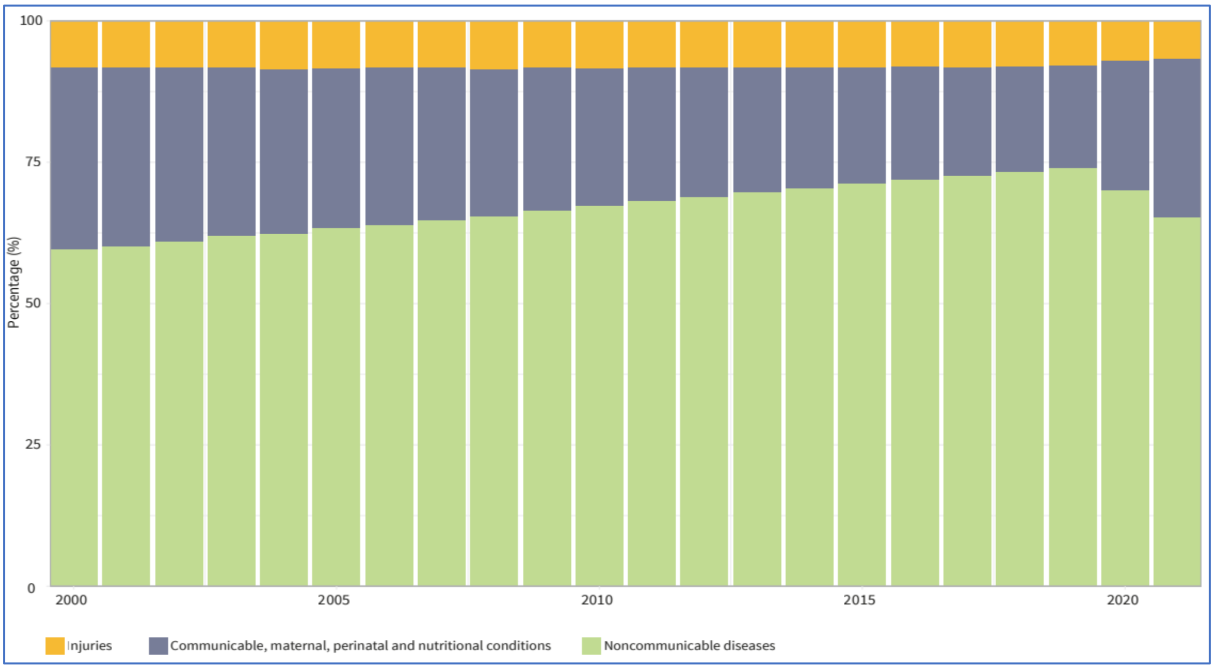


Fig 3: Communicable vs non communicable for the past twenty years

Nearly 70% of the world's fatalities occur as a result of noncommunicable diseases. These include cardiovascular disease, stroke, cancer, diabetes, and chronic obstructive pulmonary disease. There has been no discernible improvement in the fight against CVD on a global scale. Although death rates due to cardiovascular disease have been declining worldwide for the past 30 years, but this decline has stuck and become adverse if we don't take necessary measures.

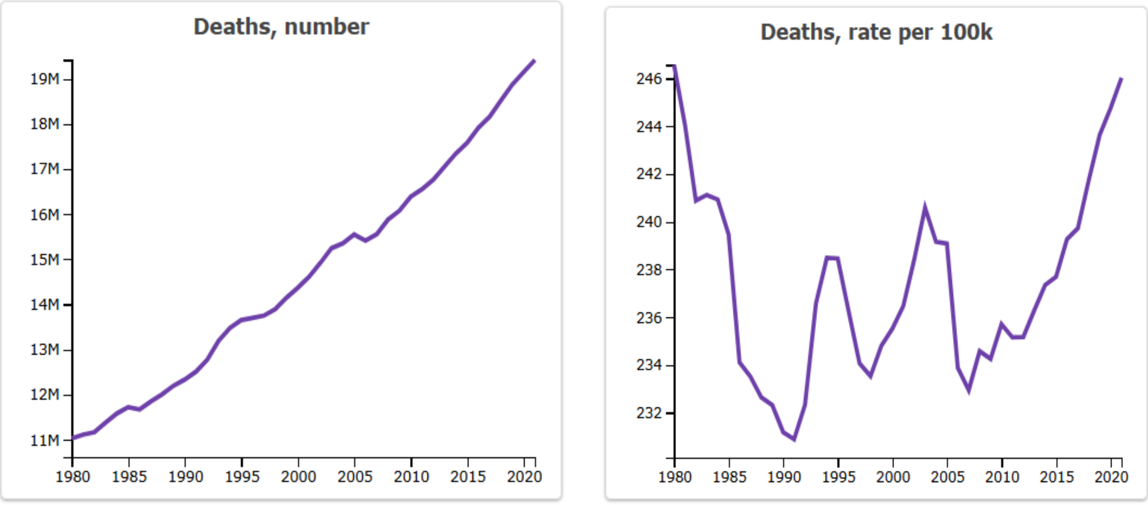


Fig 4: Number of deaths and death rate due to cardio vascular diseases during past 40 years

Socioeconomic status, metabolic factors, behavioural factors, and environmental factors are all potential causes of cardiovascular diseases (CVDs), which impact the heart or blood vessels. These encompass a wide range of health issues, such as hypertension, diabetes, heart disease, lung disease, cancer, obesity, inactive lifestyle, hazardous alcohol consumption, stress, and poor dietary habits.

Worldwide, CVDs are the prime cause of mortality rate for many years. The number of fatalities caused by cardiovascular diseases increased dramatically from 12.1 million in 1990 to 20.5 million in 2021, accounting for approximately one-third of all deaths worldwide. The leading cause of premature death for men in 146 countries and for women in 98 countries is ischaemic heart disease.

Among the cardio vascular diseases, ischaemic heart disease remains the top cause of death. In 2021, hypertension was a factor in nearly 10.8 million fatalities worldwide, making it the top risk factor for cardiovascular disease. The number of deaths and death rate due to Cardio Vascular diseases [1] [2] during the past forty years is presented in fig 4.

The second leading cause of deaths are pulmonary diseases which are also of type non communicable. Due to these two aberrant types of diseases, the death rates are at peak and the Life expectancy is coming down. Also, they account for a substantial burden worldwide. As per the WHO report, more than 8 million people are dying every year from respiratory diseases consequently in 2019 they are at the third position among the top ten killers in the world. Asthma, chronic obstructive pulmonary disease (COPD), pneumoconiosis, interstitial lung disease (ILD), and pulmonary sarcoidosis are all examples of conditions that fall under the umbrella of chronic respiratory disease (CRD).

Globally, CVDs and CRDs cause a lot of deaths and disabilities. According to the WHO, the principal coordinating body for implementing health related Sustainable Development Goals (SDG) during the period 2019-2023 stated that by 2030, the premature mortality need to be reduced by one third. About one billion plus people are targeted for better health and well-being, universal health coverage, and enhanced protection against health emergencies which are the three key goals outlined. In order to do so, the progress is to be tracked and certainly up-to-date population measurements are very much required.

The Global Burden of Diseases (GBD) [5] [6] has been measuring the impact of 369 illnesses and injuries and 87 risk factors since 1990 in 204 nations and territories which are divided into 21 regions and 7 super-regions. The leading cause of disability-adjusted life years (DALYs) and deaths worldwide among chronic respiratory diseases (CRDs) is chronic obstructive pulmonary disease (COPD). In 2019, COPD caused 3.3 million (2.9-3.6) mortalities worldwide, with 212.3 million (200.4-225.1) prevalent cases and 16.2 million (15.2-17.2) new cases.

In this introduction section, the top most health disorder influencers and their global impact are discussed. In the next section, literature on the top two global burdensome diseases is presented. In the third section, the material and methodology used in this research have been discussed while the results and discussion is presented in the fourth section. Finally, the paper is concluded in fifth section.

2. Literature study:

Heart and blood vessel diseases encompass a range of interconnected medical issues and they are the targeted elements of cardiovascular disease. The circulatory system ensures our survival by delivering oxygen, nutrients, and water to various sections of our body. It has numerous organs, all of which are vital. The lungs play a pivotal role in delivering oxygen to all of our internal organs via the blood circulation system. This blood is pumped around the body by the powerful heart muscle. Arteries and veins are the two main kinds of blood vessels.

The arterial system typically carries blood that is rich in oxygen from the heart to various organs. Instead of delivering oxygen-rich blood to our organs, veins transport carbon dioxide-rich blood back to our heart. They function well in various organs making the entire system intact, but they are most dangerous when they hit the cardiovascular system, the pulmonary system, or the central nervous system. Because of their critical role in blood circulation, the heart and lungs are particularly vulnerable to damage to their blood vessels. Also due to its high sensitivity to fluctuations in blood flow, the brain is susceptible to cardiovascular diseases.

Hypertension (high blood pressure), high cholesterol, smoking, pollutants, inflammation, and other age-related processes are among the many risk factors to which humans are vulnerable. Blood vessels can be damaged over time by these risk factors. atherosclerosis and aneurysms are two forms of vascular damage. Some cardiovascular diseases include these conditions, and they can also cause other cardiovascular diseases.

Arteries, which transport oxygen-rich blood away from the heart and to our organs, are particularly vulnerable to the buildup of lipids, cholesterol, and other substances that characterise atherosclerosis. Blood flow to organs is reduced as a result of formation of such substances, which narrows the space for blood to travel through. Aneurysms occur when the walls of blood vessels become too thin and weak, causing the vessels to bulge outward. Several factors, including atherosclerosis, injuries, and hypertension, can increase the likelihood of aneurysm development. Worldwide, ischemic heart disease—also called coronary artery disease—is the leading cause of mortality from cardiovascular causes. It occurs when the coronary arteries become narrowed, a condition commonly caused by atherosclerosis. Potential consequences of ischemic heart disease include myocardial infarction which is also called as heart attack, unstable angina, and silent ischemia.

Heart muscle cells die due to an interruption in blood flow, inflammation, or damage to the heart resulting in heart failure. Chest pain or discomfort, medically known as angina, occurs when blood supply to the heart is inadequate. The heart can become damaged or even destroyed by the constant strain of high blood pressure; a condition known as hypertensive heart disease. The heart's structure changes as a result of this condition; for instance, the heart's muscles get thicker and its capacity to pump blood becomes impaired. Uncontrolled hypertension, smoking, alcohol, unhealthy eating habits (including an excess of salt and a lack of potassium), and air pollution are all contributors to the development of hypertensive heart diseases.

Various disorders affecting the heart muscle are collectively known as cardiomyopathies. Although genetic mutations are the most common cause, other risk factors, such as alcohol consumption, can also contribute to their development. Damage to the heart's middle layer of muscle tissue can lead to myocarditis. A variety of immune system disorders, including viruses, poisons, medications, and autoimmune diseases, can cause it. One kind of inflammation that can affect the heart is endocarditis. The infection usually starts with bacteria like Staphylococcus and Streptococcus, which can attach to the heart and enter the bloodstream. As a result of rheumatic fever, rheumatic heart diseases can develop.

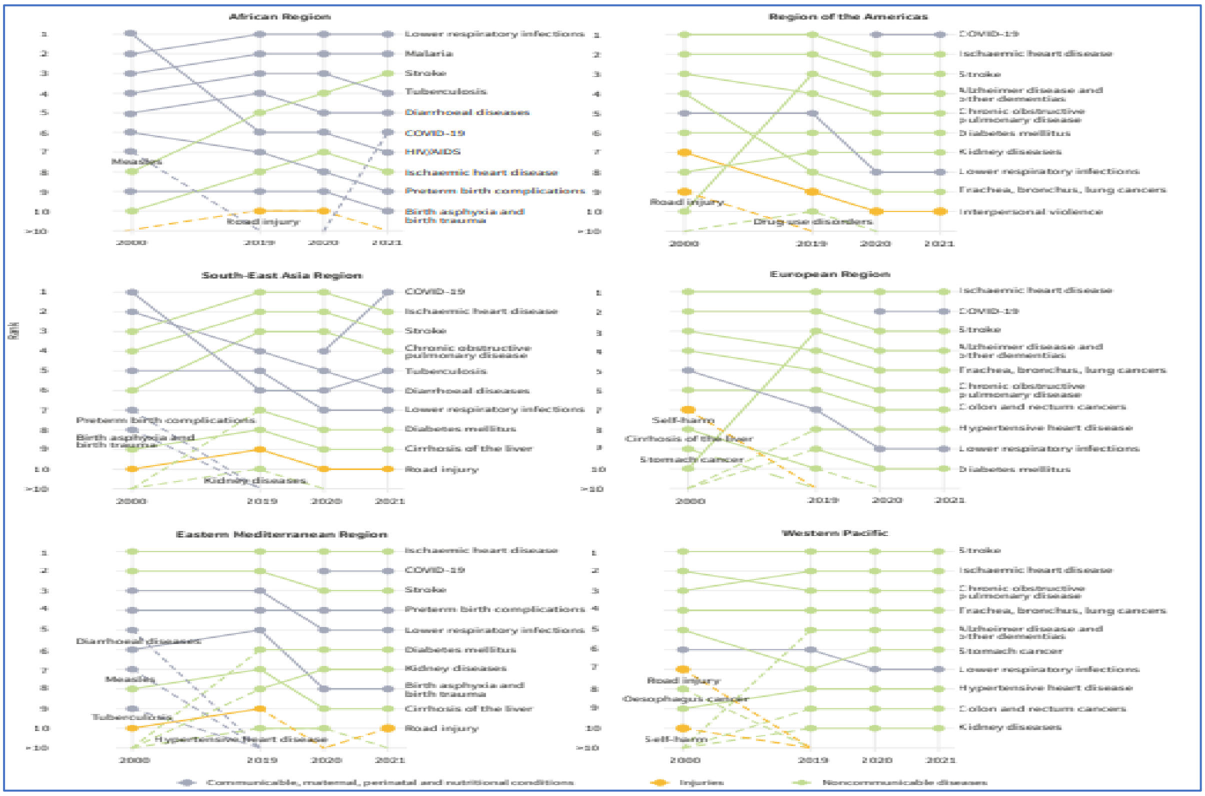


Fig 5: Region wise global mortality rate information [4]

Lungs can also be affected by diseases of the blood vessels. When blood flow to the arms and legs is reduced due to narrowed arteries, a condition known as peripheral artery disease develops. Atherosclerosis is usually the perpetrator typically affecting the legs, deep vein thrombosis (DVT) occurs when a blood clot forms in a deep vein. Another major complication of deep vein thrombosis is the possibility of the blood clot breaking off and travelling to other organs, particularly the lungs, where it could cut off blood flow. When a blood clot blocks an artery leading to the lungs, it is known as a pulmonary embolism. Deep vein thrombosis is a known risk factor for its development. This is because blood clots typically originate in other areas of the body and then make their way to the lungs, most commonly from deep veins in the legs. Pulmonary hypertension is characterised by elevated blood pressure within the pulmonary arteries, often accompanied by a dilated right ventricle in the heart. Chronic obstructive pulmonary disease (COPD) and other lung and heart conditions, as well as artery diseases, can lead to its development.

Pulmonary infections include chronic obstructive pulmonary disease (COPD), asthma, acute lower respiratory tract infection (ALRTI), tuberculosis (TB), lung cancer (LC), and other important respiratory disorders including sleep disorder breathing, pulmonary hypertension (PHN), pulmonary embolism (PE), interstitial lung disease (ILD), and bronchiectasis. Conversely, cardiovascular disorders include conditions such as coronary artery disease (CAD), arrhythmias, heart failure, heart valve disease, pericardial disease, cardiomyopathy, and congenital heart disease. Behavioural factors, like lacking in enough exercise, eating too much salt, drinking too much alcohol, and smoking. Hypertension, diabetes, obesity, high fasting plasma glucose, high levels of low-density lipoprotein (LDL) cholesterol, and high body mass index are all metabolic risk factors. Air pollution is also another factor which is categorized under environmental factors.

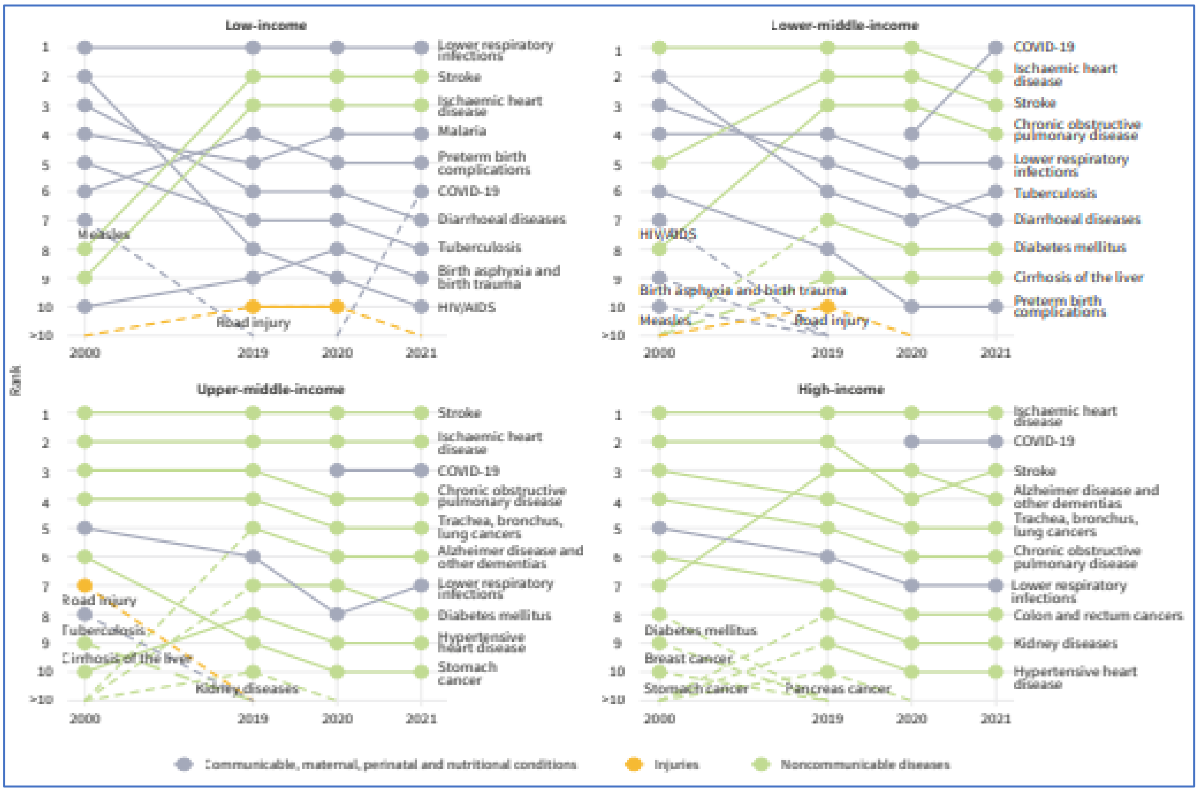


Fig 6: Income wise global mortality rate information

Further a study is done on how the heart and lung diseases are influential in geographical as well economic perspectives. Geographically in all regions except in African region, heart diseases are the highest causes of mortality. In African region, the top three reasons for mortality are lower respiratory disorders, malaria and stroke. Although the mortality in this region due to stroke was very less in 2000, it is highly progressive in the recent times. Similarly in the income perspective, heart diseases are the prime factor in all countries of various income levels. However, in low-income countries respiratory disorders are the major reason. Next to respiratory disorders are the heart diseases. The details region as well income wise are portrayed graphically in figures 5 and 6 respectively.

Having gone through the literature, perceiving the alarming situation and the dreadful diseases, the focus is on Cardio vascular and Pulmonary related disorders. However, in this paper a dataset is chosen and analysis is performed. Exploratory statistical analysis and Machine Learning mechanisms have been applied on the data and observations are cited. The succeeding section demonstrates the material and methodologies used in a detailed manner and further results are comprehensively discussed.

3. Materials and Methods:

In this section, the data analysis methods are explored and interpreted. The initial sub sections cover the statistical methods and in the later part, couple of Machine learning methods are explored. The data used for the analysis is a dataset [8] that contain around 2 lakh records. The number of features in the dataset is 21. Table 1 represents the features, description of each feature and the values that each feature can contain.

Exploratory data analysis (EDA) is a cardinal step when performing research and it aims at investigating the distributional trends, outliers, and anomalies in the data so that the hypothesis could be tested more precisely. Graphical representation enhances the Visualization and understandability of the data, which is another tool for hypothesis generation. There are quantitative methods, but most EDA approaches are more graphical. Graphics provide analysts unparallel power due to which they heavily rely upon exploratory data analysis.

EDA Technique	**Purpose**
Histogram	Distribution of a variable can be understood. Also outliers could be identified.
Scatter plots & whisker plots	Outliers can be identified
2D scatter plot, Covariance, correlation and curve fitting	Quantification of relationship between 2 variables
Heatmaps	Visualization of relationship between variables
t-SNE PCA in combination with Scatter plot	High dimensional data visualization

Table : Various EDA techniques and their purpose

SNo	**Feature**	**Description**	**Data type**	**Values**
1	USMER	Type of institution of the National Health System	int64	[2, 1]
2	MEDICAL_UNIT	Type of institution of the National Health System	int64	[1, 2, 3, 4]
3	SEX	Gender of the patient (1 = female, 2 = male)	int64	[1, 2]

4	PATIENT_TYPE	Type of care the patient received (1 = returned, 2 =)	int64	[1, 2]
5	DATE_DIED	Date of death	object	[03-05-2020, 03-06-2020, 09-06-2020, ...
6	INTUBED	Showing the different level of ventilator a patient	int64	[3, 1, 2, 4]
7	PNEUMONIA	Showing the patient have air sacs inflammation	int64	[1, 2, 99]
8	AGE	Age of the patient	int64	[65, 72, 55, 53, 68, 40, 64, 37, 25, 38, ...
9	PREGNANT	Whether the patient is pregnant or not	int64	[2, 3, 4, 1]
10	DIABETES	Whether the patient has diabetes	int64	[0, 1]
11	COPD	Indicates whether the patient has Chronic obstructive pulmonary disease	int64	[0, 1]
12	ASTHMA	Whether the patient has asthma	int64	[0, 1]
13	INMSUPR	Whether the patient is immunosuppressed	int64	[0, 1]
14	HIPERTENSION	Whether the patient has hypertension	int64	[1, 0]
15	OTHER_DISEASE	Whether the patient has other disease	int64	[0, 1]
16	CARDIOVASCULAR	Whether the patient has heart or blood vessels...	int64	[0, 1]
17	OBESITY	Whether the patient is obese	int64	[0, 1]
18	RENAL_CHRONIC	Whether the patient has chronic renal disease	int64	[0, 1]
19	TOBACCO	Whether the patient is a tobacco user	int64	[0, 1]
20	CLASSIFICATION_FINAL	Different level of COVID-19 disease	int64	[3, 5, 7, 6, 1, 2, 4]
21	ICU	Indicates whether the patient had been admitted	int64	[0, 1]

Table 1: Dataset features

The goals of exploratory data analysis (EDA) [9] [10] can be succinctly summarised as follows:

1. Enhance the understandability of database and its underlying structure
2. Visualise the possible connections (both in terms of direction and strength) between the variables that represent the level of exposure and the variables that represent the outcome.
3. Identify outliers and anomalies, which are values that differ significantly from the rest of the observations.

4. Construct concise models which is a predictive or explanatory model that functions optimally or make a choice of suitable models at the initial stage itself. Identify and generate variables that are directly applicable to clinical settings.

Sample statistics are numerical measures that describe the properties of a subset of data, known as a sample. These statistics are often used as estimates for the corresponding population. The central tendency of the data can be expressed through various characteristics such as the arithmetic mean, median, and mode. The spread of the data can be measured using metrics like variance, standard deviation, interquartile range, maximum and minimum values. Additionally, certain features of the data's distribution, such as skewness and kurtosis, can also be analysed. The arithmetic mean, also known as the mean, is calculated by dividing the sum of all data points by the total number of values. The variance σ2 is calculated by dividing the sum of squares by the population size, denoted as n. The standard deviation is defined as the positive square root of the variance.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x^i-\bar{x})^2}{(n-1)}$$

The interquartile range (IQR) is computed by considering the data points that lie between the first quartile (Q1) and the third quartile (Q3). Skewness quantifies the degree of asymmetry in a distribution. Kurtosis is a statistical measure that provides information about the extreme values (both the smallest and largest) of a distribution. Cross-tabulation is a fundamental technique in exploratory data analysis (EDA) that involves analysing the relationship between two variables without using graphs. It is an expansion of tabulation that is applicable to both categorical and quantitative data with a limited number of variables.

Covariance and correlation measures relationship degree between two random variables and indicate the extent to which they vary. x and y are the variables, n the number of data points in the sample, $\bar{x}$ the mean of the variable x and $\bar{y}$ the mean of the variable y.

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x^i-\bar{x})(y^i-\bar{y})}{(n-1)}$$

The scaled version of covariance is said to be correlation that is used to evaluate the linear association between two variables. It is computed using the following formula.

$$cor(x,y) = \frac{cov(x,y)}{s_x s_y}$$

Histogram is a bar plot showing counts versus subgroups for a variable. In this bar graph, each value represents a different frequency or proportion of cases across a given range. An easy alternative to histograms is the stem and leaf plot, which is also known as a stem plot. They display the distribution's shape in addition to the data values. Boxplots are great for showing symmetry, outliers, central tendency, and skewness, but they can mask multimodality and other data features. As boxplots use reliable statistics like median and IQR, they are a great EDA tool. A two-dimensional line plot is a graphical representation of an array's y-axis values shown at regular intervals along the x-axis.

Multivariate Graphical EDA

Multiple boxplots displayed side by side makes it easy to compare the features of different data sets. To construct a Scatterplot, two quantitative variables that are ordinal, continuous, or discrete can be used. The coordinates of each data point are associated with a specific variable. Additional variables to a maximum of five dimensions could be used, differentiating size, shape, or colour of the data points. Curve fitting is a method for estimating the strength of a correlation or how the variables change with respect to time. Minimising the sum of squared errors (SSE) between the fitted function and the data is the most common approach for curve fitting.

Heat map is another graphical representation which is a grid structured built using 2D array data. Each cell in the map has a colour that represents a value. This representation is very much useful how a variable affects another in relation to other variables is to be represented. when you need to show how one variable gets affected by other variables which make up a function.

Non-graphical EDA

In Non graphical EDA, the most important and widely used technique is Hypothesis testing which is used to prove an argument. It is done by formulating a null hypothesis which has the statement that is negation of the argument to be proved. Now the objective is to disprove the null hypothesis. Another statistical measure is the survival analysis. The plotting of a Kaplan-Meier survival function is the first step in analysing survival data, which is similar to 2 x 2 table in binary data analysis. The probability of surviving for at least a certain amount of time is the survival function and it is a function of time.

Cohen Kappa score or Kappa coefficient is another statistical measure used to quantify the categorical data and is done using inter-rater agreement computed with Cohen's Kappa Score. Although it can be modified to work with more than two raters, it is most commonly employed when there are two raters. One rater acts as the model for machine learning binary classifications, while the other acts as an observer with knowledge of the true classifications. Instead of focusing on accuracy, the primary goal of using the Cohen Kappa metric is to assess how consistent the classifications are. The possible values for the score are from -1 (unsatisfactory performance) to 1 (excellent performance).

$$\boldsymbol{k = \frac{p_0 - p_e}{1 - p_e}}$$

Where $\boldsymbol{p_0}$ is the observed agreement and $\boldsymbol{p_e}$ is the expected agreement.

Machine Learning

Machine Learning is part of Artificial Intelligence which concerns with the study of statistical techniques or algorithms that could be learnt from data and perform the tasks as per the requirement. In other words, Mathematical optimizations & statistical methods are the foundations of ML. Such techniques must learn the data patterns on their own, interpret as per the requirement and output the results which aid in decision making. Four types of Machine Learning mechanisms prevail in which certain tasks are performed by each category. In this paper ML algorithms are used in interpreting the data. Dimensionality reduction and classification are the two mechanisms applied on the data.

Dimensionality reduction is the method to represent any dataset using less number of features without any loss in the meaning. Certainly High-dimensional datasets have practical concerns like increased computation time, storage space and so on. Using dimensionality reduction we can overcome such limitations. In supervised learning, classification is a major task in identifying the data whether it belongs to certain category or not. To demonstrate Machine Learning, both these techniques are used in this paper.

4. Results and Discussion

Statistical and Machine Learning analysis provides the details that vary from basic to advanced intelligent computational level which are useful in wide variety of occasions. In the following subsections, various computations are presented that include univariate analysis, bi variate analysis, statistical measures, Dimensionality reduction, Hypothesis testing and Machine Learning computations.

4.1 Univariate Analysis

Univariate analysis is the analysis performed using single variable. The total number of patients in the dataset is 199999. A simple task is performed which displays the counts of recovered and dead patients which resulted in patients recovered count being 158592, while patients died count is 41407.

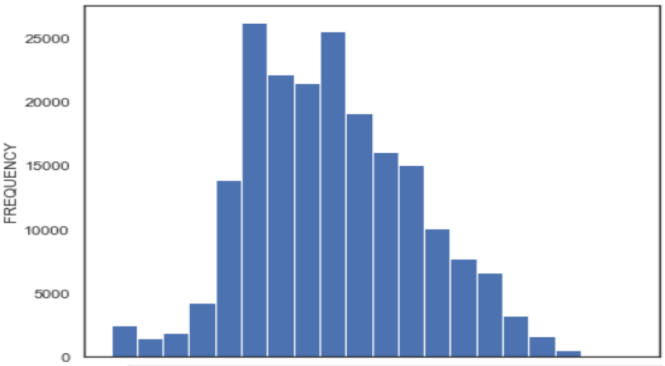


Fig 7(a): Frequency with respect to age

4.1.1 Age Distribution

Another representation is histogram analysis. The distribution of different ages and respective frequencies could be observed from figure 7(a). The same distribution is also shown in box plot in figure 7(b).

4.1.2 Statistical measures

In this subsection, the statistical measures mean, standard deviation, Inter Quartile range values of all the variables are computed and presented in table 2.

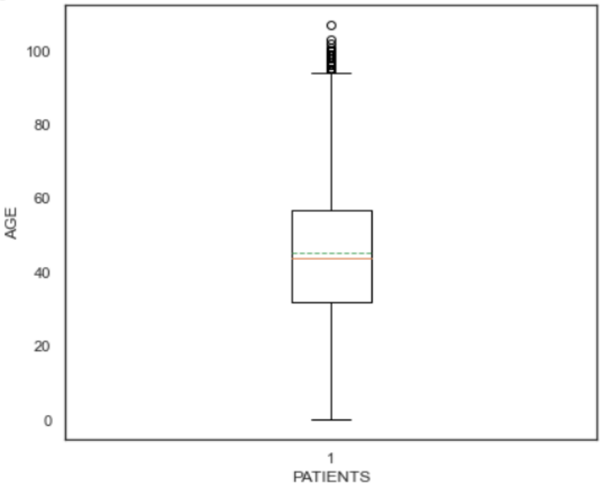
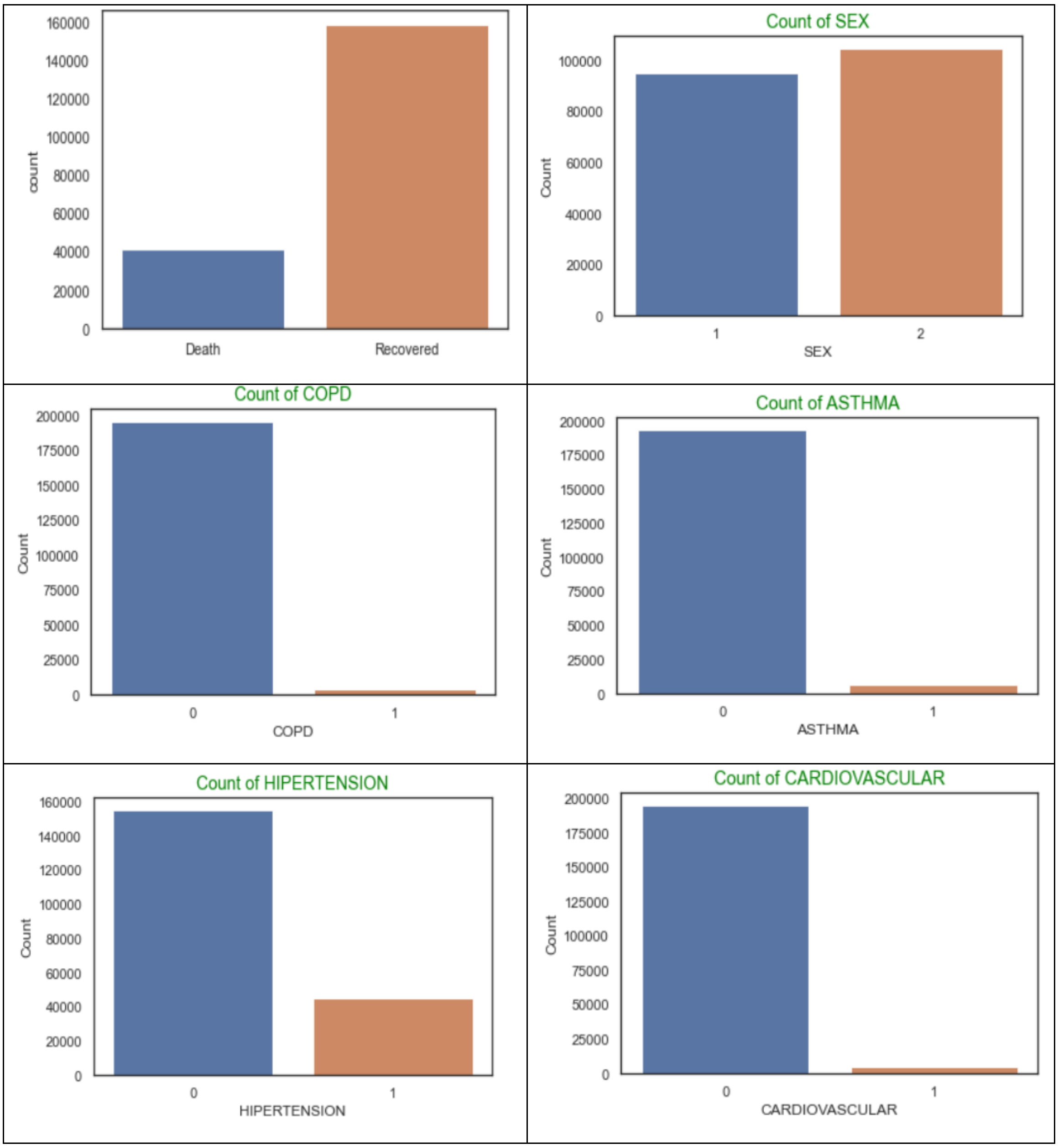


Fig 7(b): Frequency with respect to age

Feature / Variable name	count	mean	std	min	25%	50%	75%	max
USMER	199999	1.56	0.5	1	1	2	2	2
MEDICAL_UNIT	199999	3.9	0.31	1	4	4	4	4
SEX	199999	1.52	0.5	1	1	2	2	2
PATIENT_TYPE	199999	1.35	0.48	1	1	1	2	2
INTUBED	199999	2.58	0.66	1	2	3	3	4
PNEUMONIA	199999	4.1	14.82	1	2	2	2	99
AGE	199999	45.32	17.26	0	32	44	57	107
PREGNANT	199999	2.52	0.51	1	2	3	3	4
DIABETES	199999	0.17	0.38	0	0	0	0	1
COPD	199999	0.02	0.15	0	0	0	0	1
ASTHMA	199999	0.03	0.18	0	0	0	0	1
INMSUPR	199999	0.02	0.14	0	0	0	0	1
HIPERTENSION	199999	0.22	0.42	0	0	0	0	1
OTHER_DISEASE	199999	0.05	0.22	0	0	0	0	1
CARDIOVASCULAR	199999	0.03	0.16	0	0	0	0	1
OBESITY	199999	0.18	0.38	0	0	0	0	1

RENAL_CHRONIC	199999	0.03	0.18	0	0	0	0	1
TOBACCO	199999	0.08	0.27	0	0	0	0	1
CLASIFFICATION_FINAL	199999	4.01	1.52	1	3	3	6	7
ICU	199999	0.35	0.48	0	0	0	1	1

Table 2: Statistical measures of Central tendency



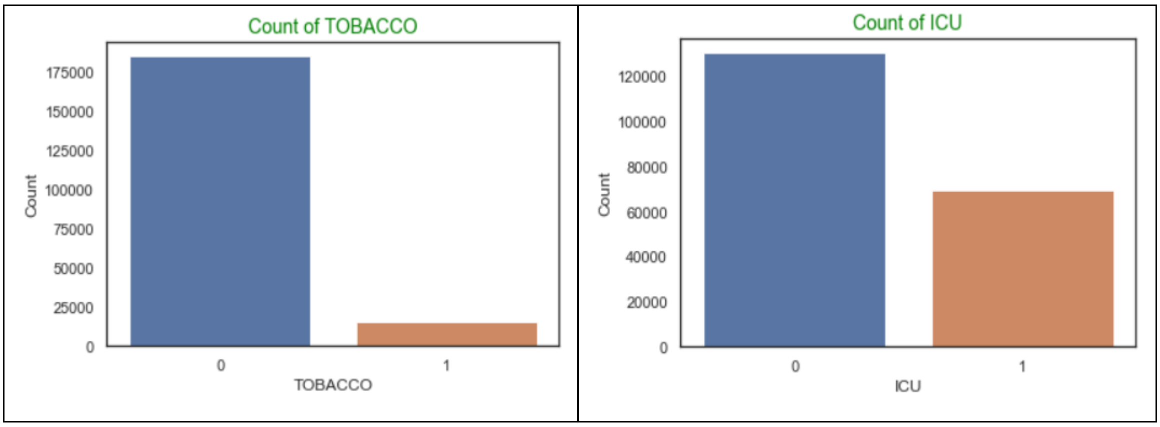


Fig 8(a) to 8(h): Patients Count with respect to various variables

Further, the patient counts with respect to different variables are presented in the form of bar charts in figures 8(a) to 8(h) respectively.

4.2 Bi-variate Analysis

Bi variate analysis involves two variables in analysing the data. On the other hand, multi variate analysis gives the information involving multiple variables. In this subsection, bi variate analysis is performed covering the following types of information.

- Gender wise count of recovered vs death. In univariate analysis simply the count of recovered vs death has been computed. Now gender wise count is computed.
- Gender wise analysis with respect to various parameters has been performed
- Survival status of patients with a disease having another disease or disorder.
- Analysis on a variable with various disorders.

4.2.1 Bi variate analysis – Gender wise count of recovered vs death

The following figure 9 depicts the counts of recovered vs non-recovered in the dataset on gender basis. The information is represented in the form of bar chart Fig 9: Representation of counts of recovered vs non recovered deaths on gender basis.as w box p

Fig 9: Representation of counts of recovered vs non recovered deaths on gender basis.

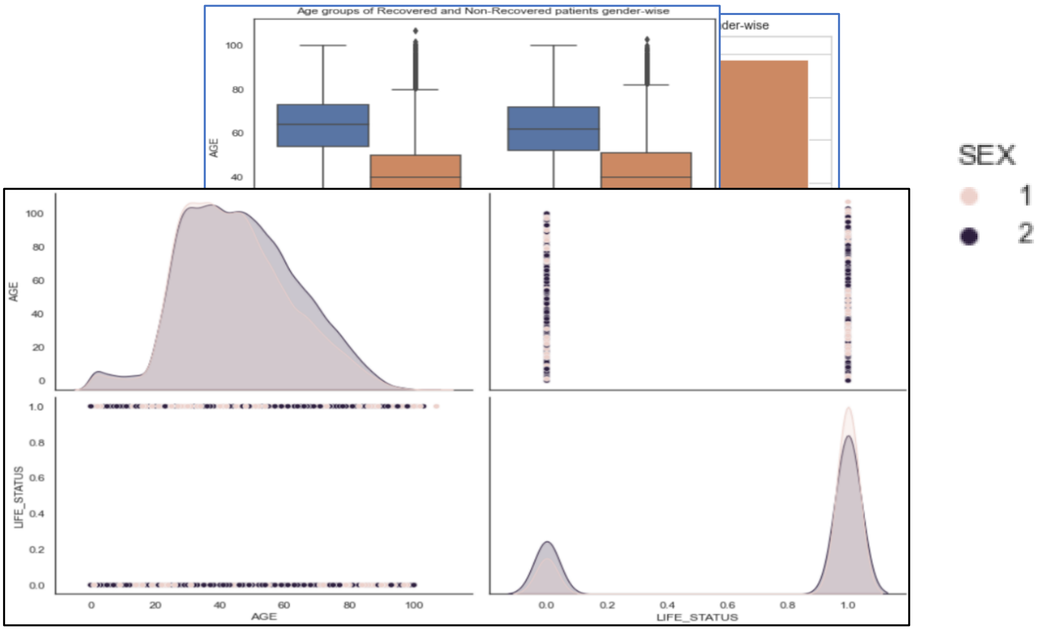
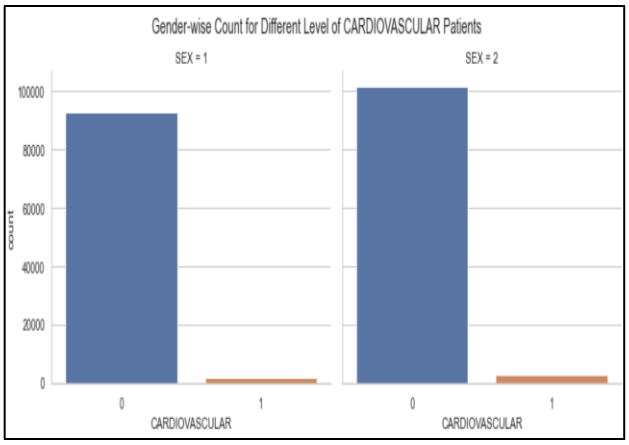


Fig 10: Representation of counts of recovered vs non recovered deaths on the basis of gender.

In figure 10 the data is represented using another graphical representation i.e., a pair plot.

4.2.2 Gender-wise Data Analysis with respect to various parameters

In bi variate analysis, another representation has been done in this section. In the below figures 11(a), (b) & (c); Gender-wise data analysis for different levels of Cardiovascular, COPD and Intubed patients has been presented.



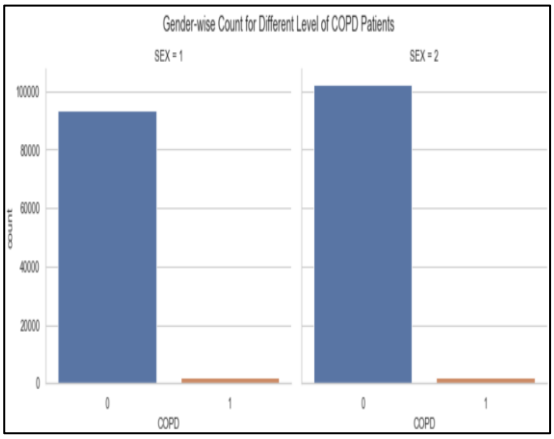


Fig 11(a): Gender-wise data analysis for different levels of Cardiovascular patients

Fig 11(b): Gender-wise data analysis for different levels of COPD patients

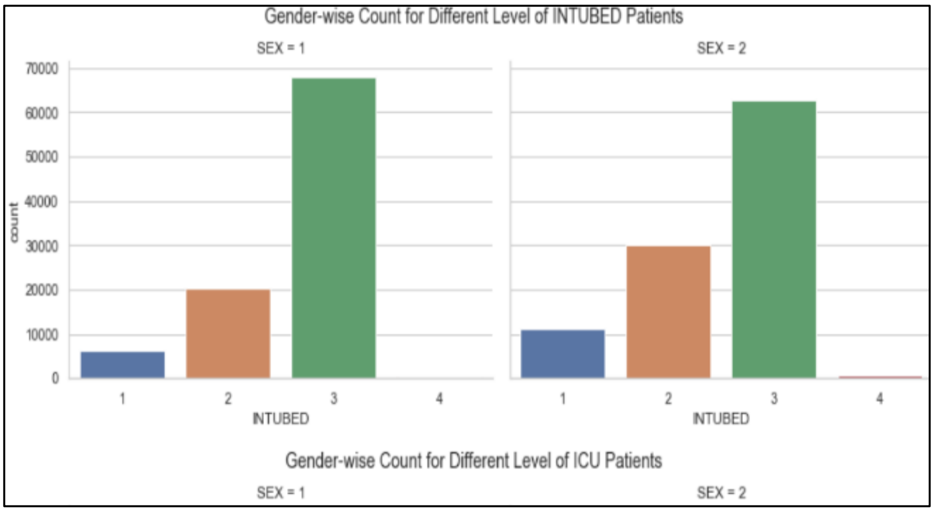


Fig 11(c): Gender-wise data analysis for different levels of and Intubed patients

From this analysis we can understand that the cardio vascular diseases are measured at two levels while the intubed patients are measured in four levels. The gender wise counts for various levels of the health disorders are computed. In providing the medicine according to the gender, the diet planning, care to be taken and the after effects of the disorders can be well planned with the data.

4.2.3 Survival status of patients with various disorders

In this subsection, under bi variate analysis, the survival status of patients with various disorders has been analyzed. These include the survival scenario of patients with tobacco, survival status of patients having hypertension along with COPD, and cardiovascular patients having COPD disorder.

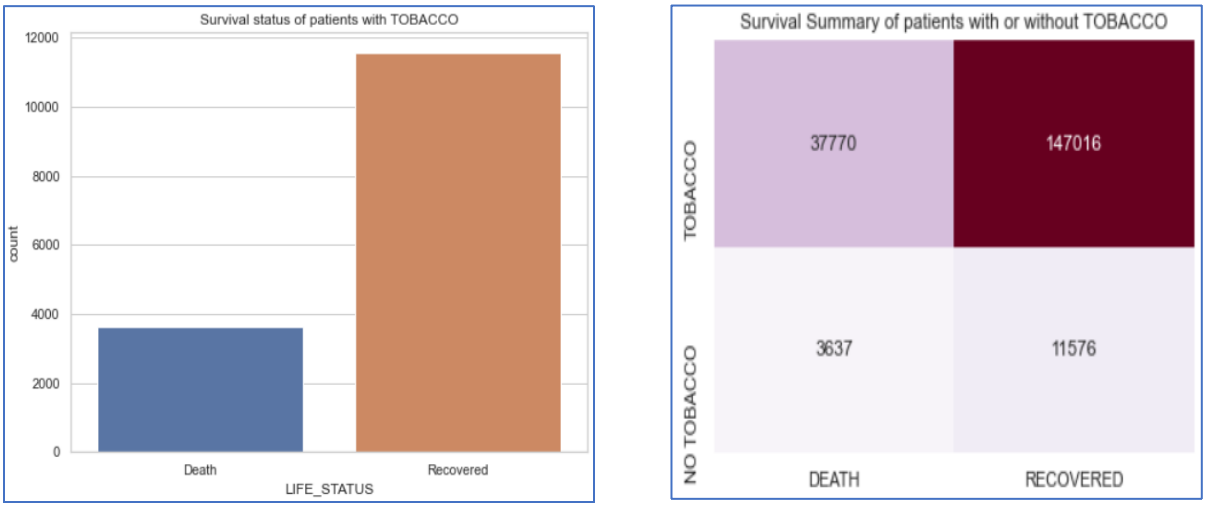


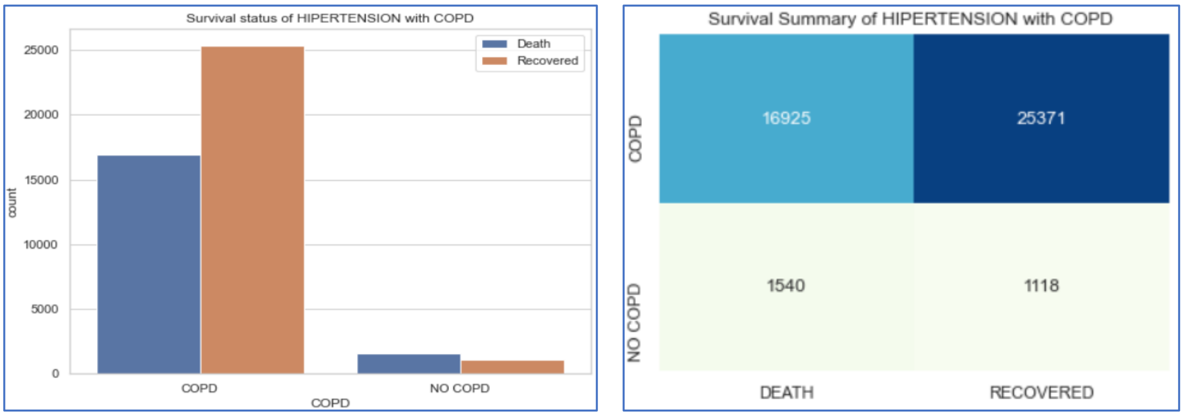
Fig 12: Survival status of patients with tobacco usage represented in bar chart & heat map.

Also, analysis has been done on whether tobacco is impacting those patients with COPD disorder which is represented in figure 12. From the above representation it could be understood that total 3637 deaths are caused due to tobacco and 11576 patients are recovered who are addicted to tobacco. On the other hand, 37770 patients survived and 147016 patients are under No Tobacco-recovered category.

Survival status of Hypertension with COPD

Figure 13 specifies that among the patients died having hyper tension the number of patients with COPD is 16925 while the number of patients not having COPD is 1540. Similarly, among the recovered number of patients, those who are having COPD is 25371 whereas 1118 patients are not having COPD.

Fig 13: Survival status of patients with hyper tension with COPD.



Survival status of Cardiovascular issues with COPD

If we observe the below figure 14, the cardiovascular with COPD patients' data is presented among which there are 2645 died patients and 2526 recovered patients.

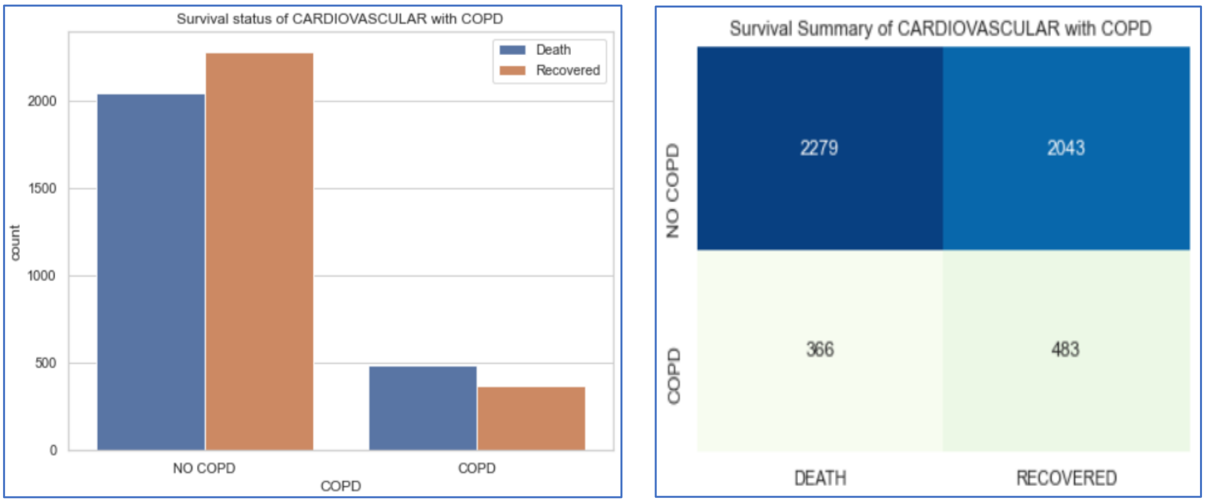


Fig 14(a): Survival status of patients with Cardio vascular with COPD

Also, it could be interpreted that there are two categories in the above data i.e., Cardio vascular patients having COPD and not having COPD. There are 4122 patients suffering with cardio vascular but not COPD and 849 with COPD.

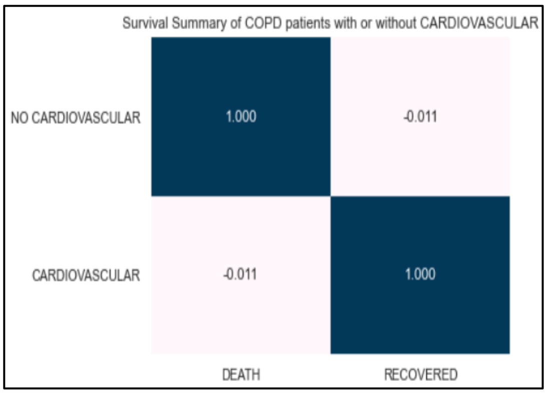
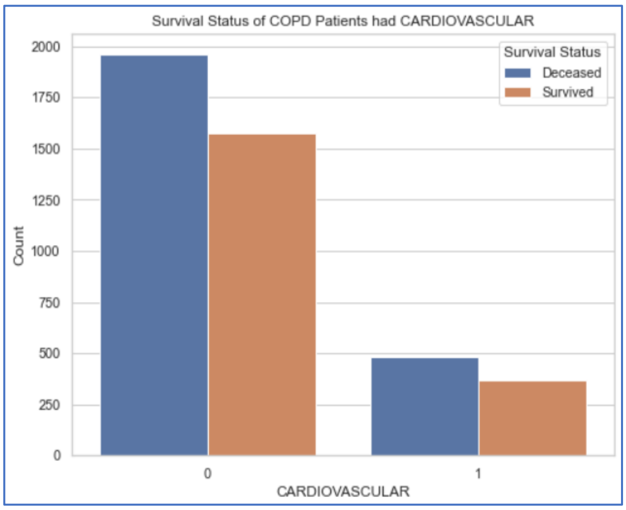


Fig 15: Survival status of patients with COPD with or without Cardio vascular disorders



However, in figure 15 the survival status of COPD patients having Cardio vascular diseases have been presented. The information is demonstrated using bar graph and heat map which specifies the correlation between the two variables.

Tobacco usage impacts COPD

In figures 13 & 14 the plots of survival status are presented. In figure 15, in addition to the bar graph, correlation heat map is presented. In figure 16, all the three representations are shown. The figures demonstrate the survival status of tobacco users with COPD.

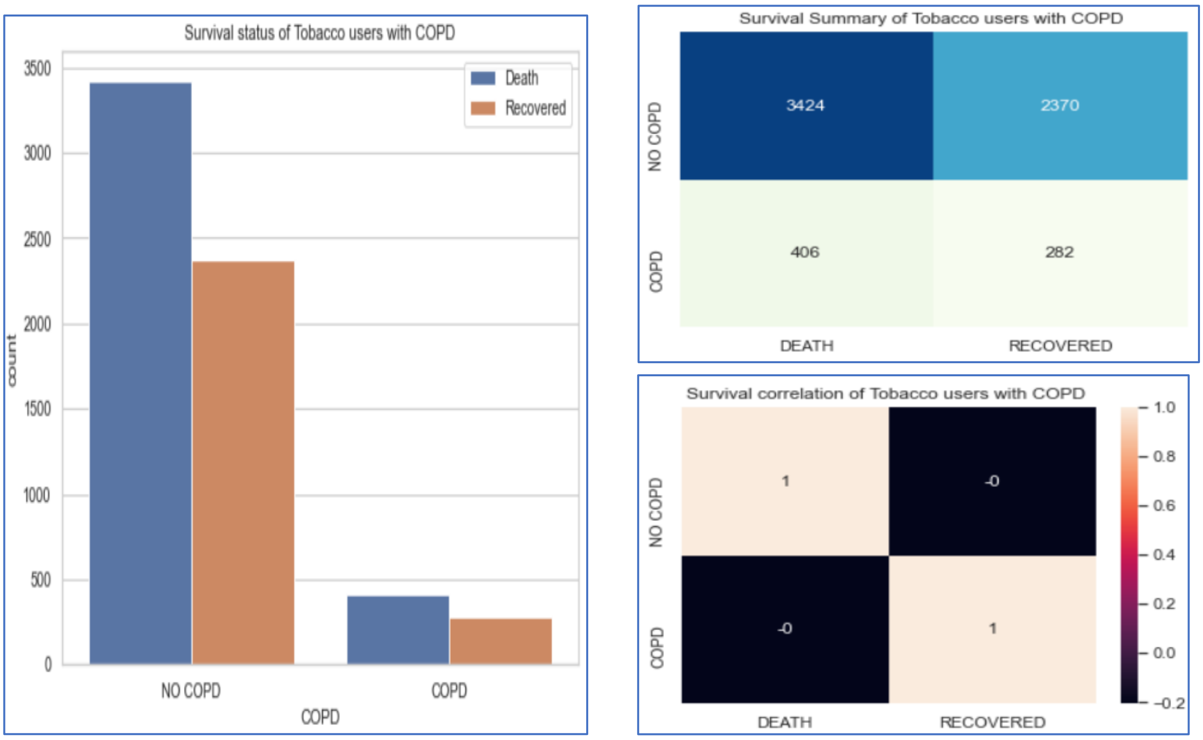


Fig 16: Survival status of tobacco users with COPD

4.3 Statistical Measures of spread depicting the age analysis with respect to few disorders:

Measures of spread describe the similarity as well as variation in the values of a dataset. Measures of central tendency doesn't focus on extreme values whereas measures of spread do account. Variance, Standard deviation, Inter Quartile Range (IQR), skewness and kurtosis are the measures of spread. Variance and standard deviation are the measures of variability. Variance indicates how the data points are away from the mean. To standardize the variance measure, Standard deviation is used and it is the root of variance. A quartile divides data into four parts at every 25 percent of values. IQR is the difference between upper and lower quartile ranges,

If the distribution of data deviates from the normal distribution it is said to have skewness. There are three types of skewness – zero skew, positively skewed and negatively skewed. In other words, skewness indicates the symmetricity of the data. If there is no any deviation, it is having zero skew while positively skewed data has the skew value greater than zero and negatively skewed data has the value less than zero. With the help of skewness we can identify outliers of the data. In positively skewed distribution, the outliers lie on the longer side right to the mean and in the other case, the outliers lie on the longer side left to the mean.

Kurtosis describes the shape of the tails of data. It doesn't speak about the peak of the distribution curve. If kurtosis is equal to 3, the tails are same as normal distribution and is also said as Meso-kurtosis. If it is greater than 3, it indicates that the tails are long and the centre is thin and tall and is called as Lepto-kurtosis and if the value of kurtosis is less than 3, it is said as Platy-kurtosis and in this case, the tails are short and the centre is broad and shorter.

Various statistical measures of spread and the distribution curve depicting the Age analysis with respect to pneumonia, COPD, asthma and Tobacco is depicted in the figure 17. The statistical measures of spread are as shown in table 3.

Statistical measure	Value
Variance	297.98
Standard Deviation	17.26
IQR	25.0
Skewness	0.238
Kurtosis	0.195

Table 3: Statistical measures of spread

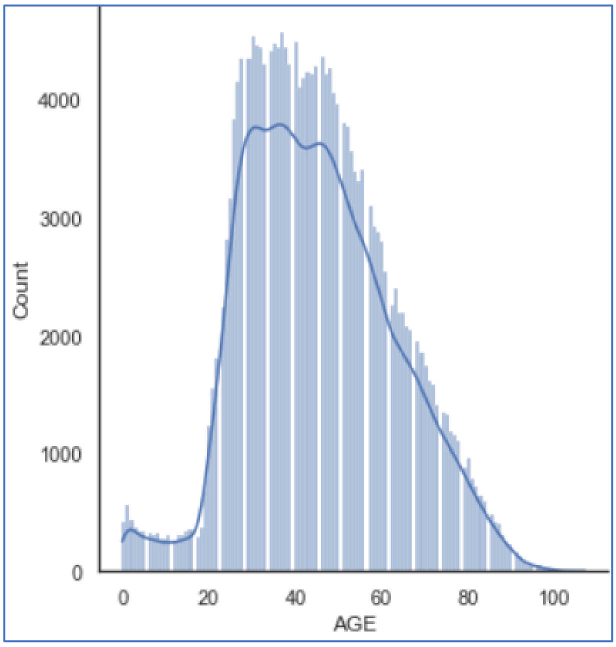


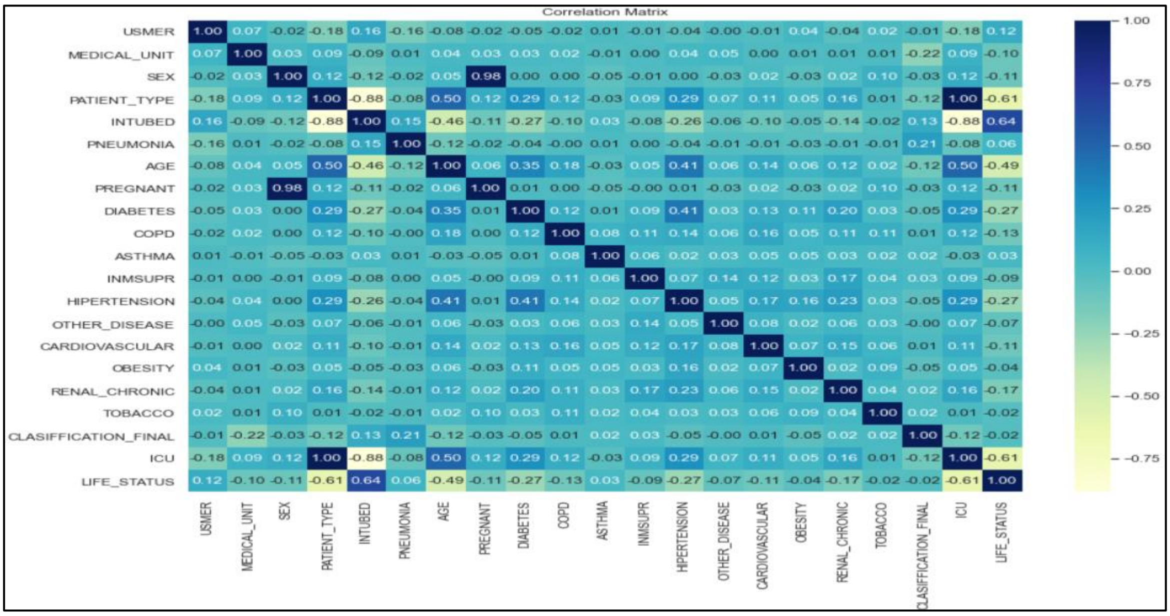
Fig 17: Spread of disorders with respect to age.

The value of variance being 297.98 seems to be larger in the dataset which means the data is having larger spread. Standard deviation indicates that the data points are at an average distance of 17.26 to the mean.

Skewness of 0.238 is positively skewed and the long tail is to the right side of the mean. Kurtosis of 0.195 indicates that the distribution is broad and short at the centre and the tails are short on either side of mean.

4.4 Correlation of variables:

In this subsection, the correlation between variables is presented. Correlation between the variables represents how they are related to one another. The correlation coefficient specifies the strength of relation between any pair of variables. Correlation matrix is generated using a method in python in which popular correlation coefficient measure called “pearson’s correlation coefficient” is inherently used. Correlation matrix for the variables in the considered dataset is as shown in figure 18. Similarly, the correlation between a single variable with other variables in the correlation matrix can also be represented. In figure 19, the correlation of different variables with "Life_Status" is presented using the bar graph. The last row in the correlation heat map in figure 18 is clearly represented below. Features those are negatively correlated as well positively correlated could be seen clearly from the graphical representation.



	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	CLASIFFICATION_FINAL	ICU	LIFE_STATUS
USMER	1.00	0.07	-0.02	-0.18	0.16	-0.16	-0.08	-0.02	-0.05	-0.02	0.01	-0.01	-0.04	-0.00	-0.01	0.04	-0.04	0.02	-0.01	-0.18	0.12
MEDICAL_UNIT	0.07	1.00	0.03	0.09	-0.09	0.01	0.04	0.03	0.03	0.02	-0.01	0.00	0.04	0.05	0.00	0.01	0.01	0.01	-0.22	0.09	-0.10
SEX	-0.02	0.03	1.00	0.12	-0.12	-0.02	0.05	0.98	0.00	0.00	-0.05	-0.01	0.00	-0.03	0.02	-0.03	0.02	0.10	-0.03	0.12	-0.11
PATIENT_TYPE	-0.18	0.09	0.12	1.00	-0.88	-0.08	0.50	0.12	0.29	0.12	-0.03	0.09	0.29	0.07	0.11	0.05	0.16	0.01	-0.12	1.00	-0.61
INTUBED	0.16	-0.09	-0.12	-0.88	1.00	0.15	-0.46	-0.11	-0.27	-0.10	0.03	-0.08	-0.26	-0.06	-0.10	-0.05	-0.14	-0.02	0.13	-0.88	0.64
PNEUMONIA	-0.16	0.01	-0.02	-0.08	0.15	1.00	-0.12	-0.02	-0.04	-0.00	0.01	0.00	-0.04	-0.01	-0.01	-0.03	-0.01	-0.01	0.21	-0.08	0.06
AGE	-0.08	0.04	0.05	0.50	-0.46	-0.12	1.00	0.06	0.35	0.18	-0.03	0.05	0.41	0.06	0.14	0.06	0.12	0.02	-0.12	0.50	-0.49
PREGNANT	-0.02	0.03	0.98	0.12	-0.11	-0.02	0.06	1.00	0.01	0.00	-0.05	-0.00	0.01	-0.03	0.02	-0.03	0.02	0.10	-0.03	0.12	-0.11
DIABETES	-0.05	0.03	0.00	0.29	-0.27	-0.04	0.35	0.01	1.00	0.12	0.01	0.09	0.41	0.03	0.13	0.11	0.20	0.03	-0.05	0.29	-0.27
COPD	-0.02	0.02	0.00	0.12	-0.10	-0.00	0.18	0.00	0.12	1.00	0.08	0.11	0.14	0.06	0.16	0.05	0.11	0.11	0.01	0.12	-0.13
ASTHMA	0.01	-0.01	-0.05	-0.03	0.03	0.01	-0.03	-0.05	0.01	0.08	1.00	0.06	0.02	0.03	0.05	0.05	0.03	0.02	0.02	-0.03	0.03
INMSUPR	-0.01	0.00	-0.01	0.09	-0.08	0.00	0.05	-0.00	0.09	0.11	0.06	1.00	0.07	0.14	0.12	0.03	0.17	0.04	0.03	0.09	-0.09
HIPERTENSION	-0.04	0.04	0.00	0.29	-0.26	-0.04	0.41	0.01	0.41	0.14	0.02	0.07	1.00	0.05	0.17	0.16	0.23	0.03	-0.05	0.29	-0.27
OTHER_DISEASE	-0.00	0.05	-0.03	0.07	-0.06	-0.01	0.06	-0.03	0.03	0.06	0.03	0.14	0.05	1.00	0.08	0.02	0.06	0.03	-0.00	0.07	-0.07
CARDIOVASCULAR	-0.01	0.00	0.02	0.11	-0.10	-0.01	0.14	0.02	0.13	0.16	0.05	0.12	0.17	0.08	1.00	0.07	0.15	0.06	0.01	0.11	-0.11
OBESITY	0.04	0.01	-0.03	0.05	-0.05	-0.03	0.06	-0.03	0.11	0.05	0.05	0.03	0.16	0.02	0.07	1.00	0.02	0.09	-0.05	0.05	-0.04
RENAL_CHRONIC	-0.04	0.01	0.02	0.16	-0.14	-0.01	0.12	0.02	0.20	0.11	0.03	0.17	0.23	0.06	0.15	0.02	1.00	0.04	0.02	0.16	-0.17
TOBACCO	0.02	0.01	0.10	0.01	-0.02	-0.01	0.02	0.10	0.03	0.11	0.02	0.04	0.03	0.03	0.06	0.09	0.04	1.00	0.02	0.01	-0.02
CLASIFFICATION_FINAL	-0.01	-0.22	-0.03	-0.12	0.13	0.21	-0.12	-0.03	-0.05	0.01	0.02	0.03	-0.05	-0.00	0.01	-0.05	0.02	0.02	1.00	-0.12	-0.02
ICU	-0.18	0.09	0.12	1.00	-0.88	-0.08	0.50	0.12	0.29	0.12	-0.03	0.09	0.29	0.07	0.11	0.05	0.16	0.01	-0.12	1.00	-0.61
LIFE_STATUS	0.12	-0.10	-0.11	-0.61	0.64	0.06	-0.49	-0.11	-0.27	-0.13	0.03	-0.09	-0.27	-0.07	-0.11	-0.04	-0.17	-0.02	-0.02	-0.61	1.00

Fig 18: Correlation matrix of each feature with one another

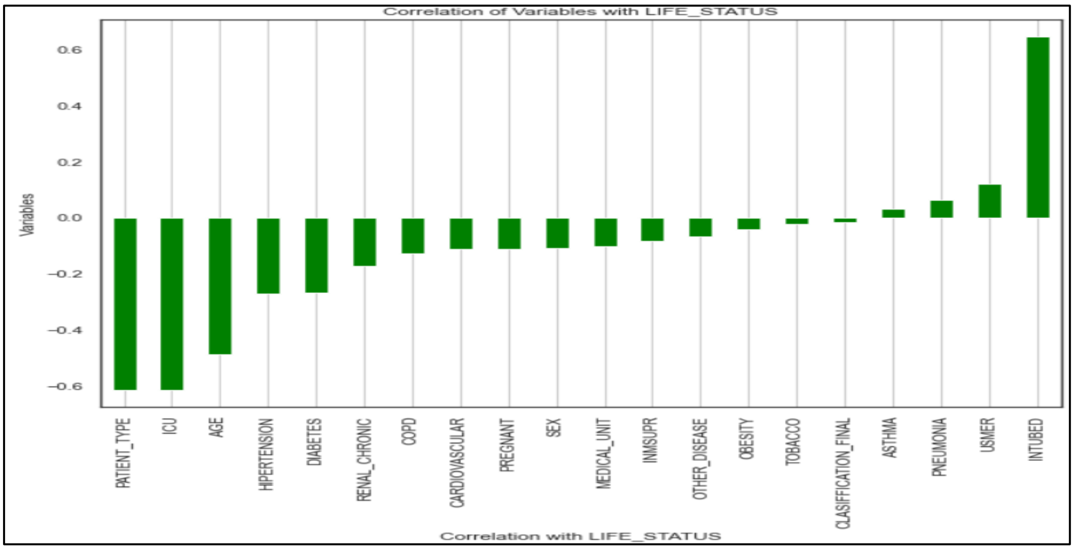


Fig 19: Correlation representation of Life status with all other variables

4.5 Hypothesis Testing:

Hypothesis Testing is a significant mechanism used in statistical analysis with which we can prove the required statements through proper analysis. It tests an experiment or sample against the population parameter. A statement about a population parameter is said to be a hypothesis. Usually we have two hypotheses – Null hypothesis (H0) and Alternate Hypothesis (H1). The null hypothesis is that it is assumed to be correct while the alternate hypothesis is that which describes what happens if the null hypothesis is incorrect. In this subsection the two hypothesis tests carried out are presented.

Case 1:

Null Hypothesis (H0):

Tobacco usage will no impact COPD.

There is a significant association between intubation (ventilator) and ICU for COPD patients.

Alternative Hypothesis (H1):

Tobacco usage will impact COPD.

There is a significant association between intubation (ventilator) and ICU for COPD patients.

Case 2:

Null Hypothesis (H0):

There is no significant association between the following pairs of categorical variables:

Alternative Hypothesis (H1):

There is a significant association between the pairs of categorical variables

In hypothesis testing, to reject the null hypothesis, chi square test for goodness of fit is conducted. A chi square test is conducted only if the conditions are met.

i) The sample method must be random.
ii) The variable must be of type categorical
iii) For each category the expected value of observations must be at least five
iv) The variable under study needs to be independent.

Another statistical measurement used in Hypothesis testing is p-value. It is used in validating a hypothesis against observed data. A p-value is a probability measure that obtains the probability of observed results presuming the null hypothesis to be true. These words kept simple, if the probability of observations in case of null hypothesis being true is very low, then null hypothesis fails and alternative hypothesis becomes true. If the p-value is 0.05 or less, then the statistical significance is high.

4.5.1 Case 1:

Test performed	**Variable**	**Value or outcome**
Chi-square statistic	COPD and SURVIVAL_STATUS of TOBACCO USERS	58.73
p-value	COPD and SURVIVAL_STATUS	1.80604 e-14
Chi-square statistic	ICU and SURVIVAL_STATUS	83,76
p-value	ICU and SURVIVAL_STATUS	5.56384 e-20
Chi-square statistic	Intubed and SURVIVAL_STATUS	119.91
p-value	Intubed and SURVIVAL_STATUS	8.02951 e-26

Table 4: Chi-square and p-value for different variables under consideration

The Chi-square statistic and p-value for the variables considered in this case are as shown in table 4. The chi-square value of COPD vs Survival status of Tobacco users is 58.73 and p-value in this case is 1.80604 e-14.

Test performed	**Variable**	**Value or outcome**

Chi-square statistic	INTUBED and COPD	261.17
p-value	INTUBED and COPD	2.50801 e-56
Chi-square statistic	INTUBED and CARDIOVASCULAR	186.49
p-value	INTUBED and CARDIOVASCULAR	3.48079 e-40
Chi-square statistic	COPD and CARDIOVASCULAR	342.4
p-value	COPD and CARDIOVASCULAR	1.91448 e-76
Chi-square statistic	HIPERTENSION and CARDIOVASCULAR	528.96
p-value	HIPERTENSION and CARDIOVASCULAR	4.73201 e-117
Chi-square statistic	CARDIOVASCULAR and HIPERTENSION	528.96
p-value	CARDIOVASCULAR and HIPERTENSION	4.73201 e-117
Chi-square statistic	CARDIOVASCULAR and TOBACCO	66.97
p-value	CARDIOVASCULAR and TOBACCO	2.75333 e-16
Chi-square statistic	CARDIOVASCULAR and ICU	183.99
p-value	CARDIOVASCULAR and ICU	6.49522 e-42
Chi-square statistic	TOBACCO and CARDIOVASCULAR	66.97
p-value	TOBACCO and CARDIOVASCULAR	2.75333 e-16
Chi-square statistic	ICU and COPD	250.48
p-value	ICU and COPD	2.03256 e-56

Table 5: Chi-square and p-value for different categorical variables

Both values are indicating the statistical significance of rejecting null hypothesis which means accepting alternate hypothesis. It clearly demonstrates that Tobacco usage will impacts COPD and survival status. Also, the intubation and ICU for COPD patients are significantly associated.

4.5.2 Case 2:

In table 5, the values of chi-square test and p-value of different pairs of the categorical variables have been presented. All pairs of variables are giving a result of high Chi-square value and least p-values which means that all the categorical variables are significantly associated. Hence the null hypothesis is rejected stating the alternate hypothesis is accepted.

4.6 Machine Learning analysis:

Analysis using Machine Learning also plays a vital role in discovering important aspects such as making predictions, automated learning, clustering, classification and other associated possibilities. To illustrate, dimensionality reduction and classification techniques have been applied on the covid dataset in this paper and are explained.

4.6.1 Dimensionality Reduction & Feature selection

Dimensionality reduction techniques [11] such as PCA, LDA and t-SNE enhance the performance of ML models. They preserve essential features of complex data sets by reducing the number predictor variables for increased generalizability. In this subsection, the dimensionality reduction applied on the dataset is described. Principal component analysis (PCA) is the most common dimensionality reduction method. It is a form of feature extraction, which means it combines and transforms the dataset's original features to produce new features, called principal components. PCA selects a subset of variables from a model that comprise the majority or all of the variance present in the original set of variables. PCA then projects data onto a new space defined by this subset of variables.

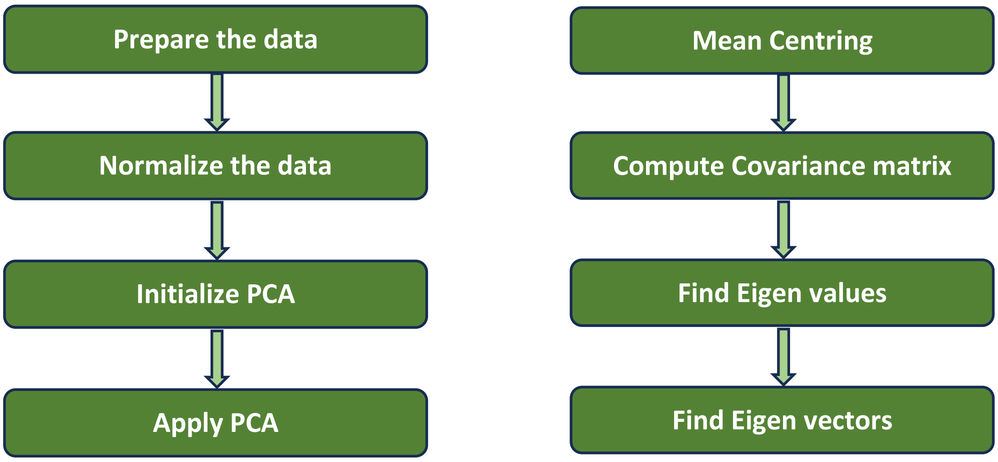


Fig 20 (a): Dimensionality reduction process **Fig 20 (b): Internal process of Dim. Red.**

The general flow of Dimensionality reduction technique is represented in figures 20(a) & (b). Outline of the technique is represented in figure (a), while the internal procedure of PCA is represented in figure (b). In the current context, the variables which are to be chosen and those variables to be dropped is part of data preparation. As a next step, standard scaler is applied. Next the PCA is initialized and applied.

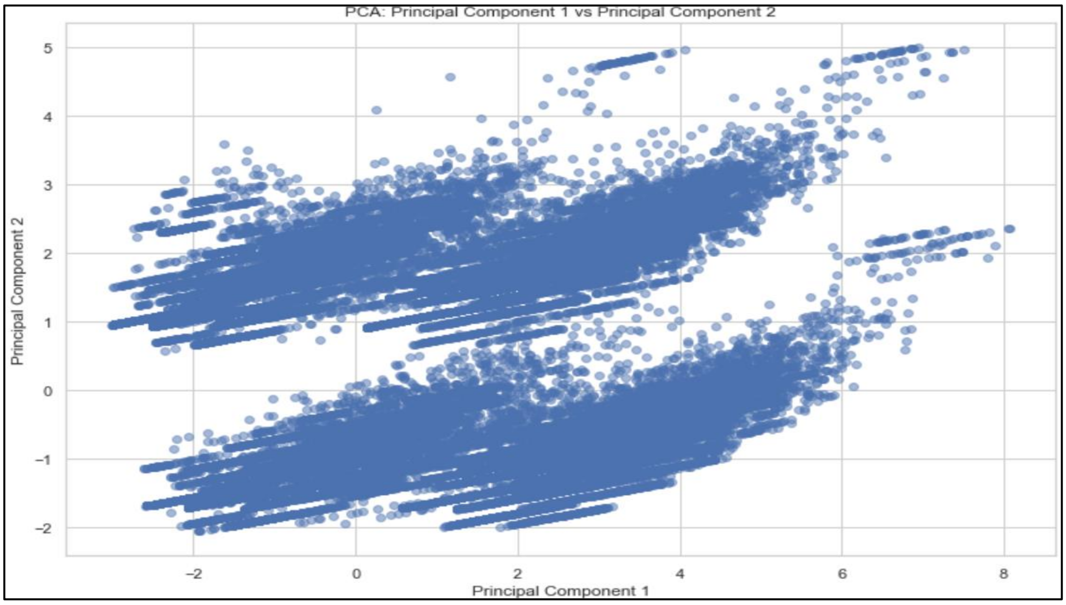


Fig 21: Principal Component loadings

In the process of applying PCA, the principal components will be selected. Depending on these components, Mean centering, covariance matrix computation happens after which the eigen values and eigen vectors are computed. In this experimentation, two principal components are chosen and the data loadings are captured in a scatter plot which are shown in figure 21. Principal component loadings mean the data is shifted to new dimensional space which represents the dimensionality reduction.

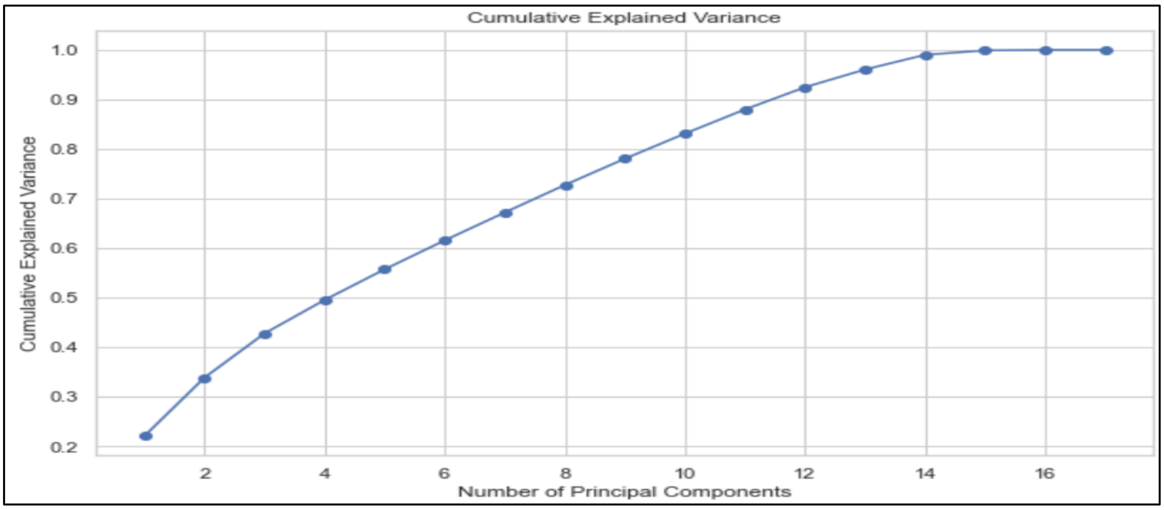


Fig 22: Cumulative explained variance for various principal components

The cumulative explained variance with various number of principal components is depicted in figure 22. Thumb rule is that the Cumulative Explained Variance crossing 90% indicates the number of Principal components required is optimum. As in the figure it is depicted to be 12, the number of dimensions can be reduced to 12 which offers better performance even without any data loss.

4.6.2 Machine Learning Modeling

In this subsection, Machine Learning has been explored. Amongst different machine learning tasks [12], classification has been applied on the data using various classifiers. In addition to the basic classification techniques, ensemble techniques are also applied. Decision Tree, K-Nearest Neighbour (KNN), Random Forest and GridSearchCV Classifiers are the techniques used and their respective metrics are tabulated in table 6.

Technique	**Train Score (%)**	**Test Score (%)**	**MSE**	**RMSE**	**R2 Score**	**Cohen Kappa Score**
Decision Tree Classifier	67.16	41.20	0.1244	0.3527	0.2388	0.614
K Nearest Neighbours	89.36	88.38	0.1161	0.3408	0.2894	0.636
Ensemble Technique (Random Forest Classifier)	91.64	87.74	0.1225	0.3500	0.2505	0.614
GridSearchCV (Random Forest Classifier)	91.90	88.03	0.1196	0.3459	0.2680	0.623

Table 6: Performance measures of various classifiers

The above results indicate that among the four techniques used, the test accuracy score is high for KNN classifier while the train accuracy is higher for GridSearchCV. On the whole GridSearchCV could be said as the top performer. It is due to the reason that, in GridSearchCV, the model has been trained well when compared to other ones and the test score is slightly less than the KNN classifier. On the other hand, Mean Square Error and the Root Mean Square Error metrics also signify the same. R2 score which is said to be coefficient of determination is a statistical measure in machine learning how well a model fits the data assessing the variance proportion in the dependent variable which could be explained by the independent variable.

Lastly, Cohen Kappa score is used to measure how well the inter-rater reliability is. The interpretation is as follows. The CK score in the range of 0.61 – 0.80 for all techniques which means the interpretation is substantial agreement. Moreover, among the techniques, the score is high for KNN classifier and next comes GridSearch CV.

With Cohen Kappa Score it could be understood that the interpretation of various algorithms is said to be rightly done. The train & test accuracies for the techniques chosen are represented in figure 23.

CKS	**Interpretation**
≤ 0	No agreement
0.01–0.20	None to slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

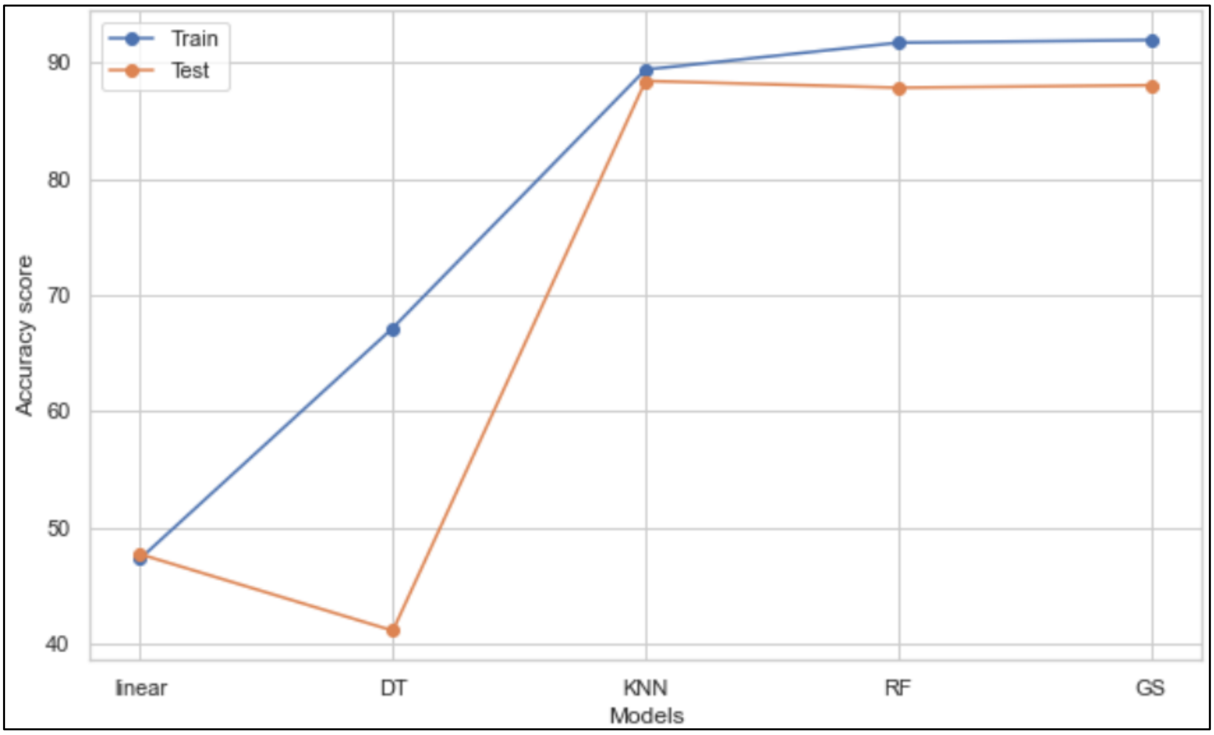


Fig 23: Train and test accuracies of various machine learning algorithms.

5. Conclusion

Countering any health disorder requires the complete information with which effective decision making could be done. Subsequently formulating and implementing the policies would also be efficient. We could do this when proper analysis and results are available. In this paper, a dataset is considered on which an exploratory analysis is done on cardio as well pulmonary disorders. The analysis encompasses various statistical methods and also machine learning methods.

Statistical methods include univariate analysis, bi variate analysis, statistical measures of spread, correlation of variables, hypothesis testing. On the other hand, dimensionality reduction and unsupervised learning are the machine learning techniques explored.

References

[1] World-Heart-Report-2023.pdf (world-heart-federation.org)

[2] Mensah GA, Fuster V, Murray CJL, et al. **Global burden of cardiovascular diseases and risks, 1990-2022.** *Journal of the American College of Cardiology.* 11 December 2023. doi: 10.1016/j.jacc.2023.11.007.

[3] World health statistics 2024: monitoring health for the SDGs, sustainable development goals (who.int)

[4] GBD 2019 Chronic Respiratory Diseases Collaborators. Global burden of chronic respiratory diseases and risk factors, 1990-2019: an update from the Global Burden of Disease Study 2019. EClinicalMedicine. 2023 May;59:101936. doi: 10.1016/j.eclinm.2023.101936. PMID: 37229504; PMCID: PMC7614570.

[5] Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet 2020; 396(10258): 1204-22.

[6] Murray CJL, Aravkin AY, Zheng P, et al. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet 2020; 396(10258): 1223-49.

[7] VizHub - GBD Results (healthdata.org)

[8] https://www.kaggle.com/datasets/somin69/covid19-dataset

[9] Komorowski, Matthieu & Marshall, Dominic & Salciccioli, Justin & Crutain, Yves. (2016). Exploratory Data Analysis. 10.1007/978-3-319-43742-2_15.

[10] NIST/SEMATECH e-Handbook of Statistical Methods

[11] Jia, W., Sun, M., Lian, J. *et al.* Feature dimensionality reduction: a review. *Complex Intell. Syst.* **8**, 2663–2693 (2022). https://doi.org/10.1007/s40747-021-00637-x

[12] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* **2**, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

[13] Yousefzadeh M, Zolghadri M, Hasanpour M, Salimi F, Jafari R, Vaziri Bozorg M, Haseli S, Mahmoudi Aqeel Abadi A, Naseri S, Ay M, Nazem-Zadeh MR. Statistical analysis of COVID-19 infection severity in lung lobes from chest CT. Inform Med Unlocked. 2022;30:100935. doi: 10.1016/j.imu.2022.100935. Epub 2022 Apr 1. PMID: 35382230; PMCID: PMC8970623.