REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

PREDICTING THE LIKELIHOOD OF A PATIENT'S SURVIVAL AFTER AN ACCIDENT USING ARTIFICIAL INTELLIGENCE AND STATISTICAL METHODS

Najwa ALSHEIKH

Master's Thesis

Supervisor

Asst. Prof. Dr. Mesut ÇEVİK

Istanbul, 2022

PREDICTING THE LIKELIHOOD OF A PATIENT'S SURVIVAL AFTER AN ACCIDENT USING ARTIFICIAL INTELLIGENCE AND STATISTICAL METHODS

Najwa ALSHEIKH

Electrical and Computer Engineering

Master's Thesis

ALTINBAŞ UNIVERSITY

2022

The thesis titled PREDICTING THE LIKELIHOOD OF A PATIENT'S SURVIVAL AFTER AN ACCIDENT USING ARTIFICIAL INTELLIGENCE AND STATISTICAL METHODS prepared by NAJWA ALSHEIKH and submitted on 21/12/2023 has been **accepted unanimously** for the degree of Master of Science in Electrical and Computer Engineering.

Asst. Prof. Dr. Mesut ÇEVİK

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Mesut ÇEVİK	Department of Electric and Electronic Engineering, Altınbaş University	_____
Asst. Prof. Dr. Timur İNAN	Department of Software Engineering, Altınbaş University	_____
Asst. Prof. Dr. Esra SAATÇI	Department of Electric and Electronic Engineering, İstanbul Kültür University	_____

I hereby declare that this thesis meets all format and submission requirements of a (Master's) thesis.
Submission date of the thesis to Institute of Graduate Studies: ___/___/___

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Najwa ALSHEIKH

Signature

DEDICATION

My family, friends, and coworkers will receive a copy of my dissertation as a token of my appreciation. My profound gratitude goes out to my wonderful parents, whose never-ending support and cheerleading motivated me to put in a lot of effort and keep going even when things became tough. My siblings have been there for me no matter what, and I hold them in the highest regard.

In addition, I would want to express my gratitude to all of my amazing friends who have supported me during the process of completing my dissertation. What they have done for me is something I will never forget.

ABSTRACT

PREDICTING THE LIKELIHOOD OF A PATIENT'S SURVIVAL AFTER AN ACCIDENT USING ARTIFICIAL INTELLIGENCE AND STATISTICAL METHODS

Alsheikh, Najwa

M.Sc., Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Dr. Mesut ÇEVİK

Date: December / 2022

Pages: (70)

The goal of this research is to construct a model that, with the help of regression methods and deep learning, will be able to compute the probability of survival for patients who are undergoing medical care in hospitals at the present time and who are exhibiting symptoms that are typically seen in intensive care units. These patients are considered to be in a high-risk category (ICUs). Survival models are a type of statistical tool that are used relatively frequently in the field of clinical research. Clinical researchers are interested in determining how long a patient is expected to live. Researchers in clinical settings may make use of survival models to the purpose of these models is to forecast the risk of a variety of clinical outcomes as well as to identify relevant risk variables. A good illustration of this would be the overall survival rate of patients who suffer from a variety of diseases. According to the findings, our model is capable of performing better than the CPH prediction model when it comes to projecting the prognosis of patients who have CCU. This is the case when it comes to the prognosis of patients who have CCU.

Keywords: ML, AI, DL, Survival Rate.

TABLE OF CONTENTS

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

APC	:	Age Period Cohort Model
MSE	:	Mean Square Error
CCU	:	Critical Care Unit
CPH	:	Cox Proportional Hazards Model
HMD	:	Human Mortality Database
LC	:	Lee Carter Model
LM	:	Lee Miller Mode
HU	:	Hyndman Ullah Model
AI	:	Artificial Intelligence
BMS	:	Booth Maindonald Smith Model
HSD	:	Human Survival Database
MR	:	Survival Rate
SSA	:	Sub-Saharan Africa
SDG-3	:	Third Sustainable Development Goal
DL	:	Deep Learning
CBD	:	Cairns-Blake-Dowd Model
LDA	:	Linear Discriminant Analysis
IMR	:	Infant Mortality Rate
TBI	:	Accidental Brain Injuries

TTAS : Taiwan Triage and Acuity Scales

GCS : Glasgow Coma Scales

CT : Computed Tomography

LR : Logistic Regression

RF : Random Forest

SVM : Support Vector Machines

MLP : Multilayer Perceptions

CART : Classification And Regression Tree

ID3 : Iterative Dichotomizer 3

ANN : Artificial Neural Networks

T-SNE : T -Distributed Stochastic Neighbor Embedding

MIT : Massachusetts Institute of Technology

OMDB : Organic Materials Database

ASM : American Society for Metals

OQMD : Open Quantum Materials Database

1. INTRODUCTION

1.1 BACKGROUND

Since the 1980s, data storage capacity has been steadily increasing, and traditional analytical tools are not powerful enough to fully exploit the value of Big Data. A large volume of data does not always lead to useful information, which makes obtaining comprehensive analyzes increasingly difficult. Therefore, analysts are turned to what is called “Machine Learning” or “Statistical Learning” or “Machine Learning”. One of the first definitions of "Machine Learning" was proposed in 1959 by American electrical and computer engineer Arthur Samuel, who defined it as a field of study that gives computers the ability to learn without being explicitly programmed [1].

Machine learning has allowed humans to not only improve many industrial and professional processes but also to advance everyday life. Virtual personal assistants use a machine learning algorithm, like Siri, Alexa, and Google. They help to find information when asked by voice, recall our corresponding queries or send a command to other resources such as phone applications, to fulfill our requests. Machine learning is also used in navigation services to predict traffic. In addition, social networks use it, for example, to personalize the news feed and target ads. Take the example of Facebook, it takes into account the friends with whom users connect, the profiles consulted most often, their centers of interest, etc. Based on continuous learning, a list of uses will be suggested.

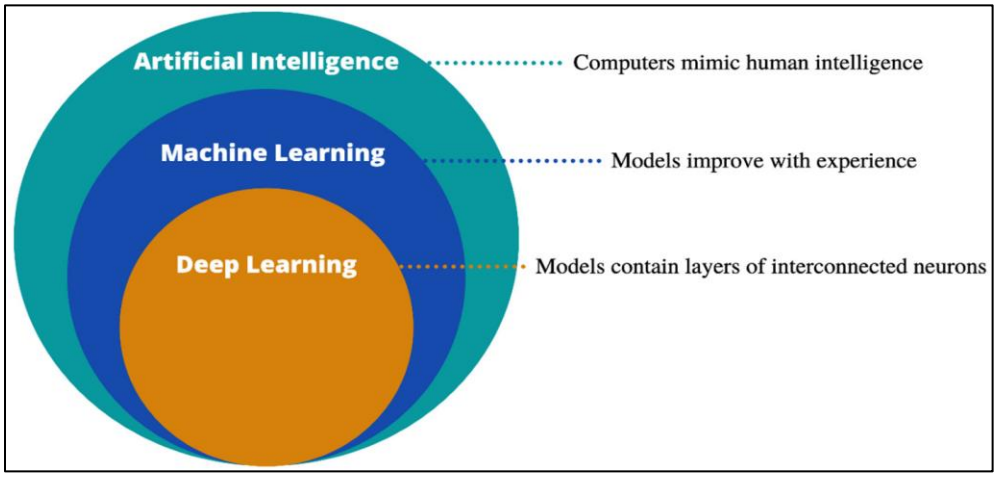

Figure 1.1: Three concepts of AI [1].

Facebook also uses face recognition based on machine learning, which is also adapted in surveillance cameras. This technology is also used by search engines like Google to refine results; with each search performed, algorithms at the server level monitor how you respond to the results. If you stay on a web page for a long time, the search engine assumes that the results displayed correspond to the query [2].

The principle of Machine Learning (ML) consists in giving computers input values to analyze and the result that we want to achieve so that the machine can create a model capable of deducing new information from old data. Analysts aim to obtain a specific program that can be executed by a computer in order to solve a problem or a well-defined task. Machine Learning users called "Data Scientists" do not try to write a program on their own, instead they collect the input data and desired target values. Then they tell the computer to find a program (or algorithm) that calculates an output for each added input value. ML is applied to complicated problems that humans are unable to solve. But the Data Scientist must always be ready and able to test a few approaches before choosing a satisfactory approach to obtain a better model. Machine learning can be applied in two ways. The first, by supervised learning which uses algorithms based on input data labeled with the desired outputs; if we have a set of objects and for each object an associated target value, we must learn a model capable of predicting the correct target value of a new unlabeled object. The second, unsupervised learning where the data is unlabeled, i.e., it has no target value.

The algorithm in this case finds on its own common points between its input data, then it classifies them into groups of data as distinct as possible by maximizing their homogeneity; each group has data with similar characteristics. The greater the number of input data, the closer we get to finding a better classification or grouping by choosing the right characteristics [2]. Materials science uses numerical methods to manage the complexity of problems in all areas of technology. Researchers in academia and industry are looking for new, high-quality materials to meet the needs of specific applications. In this context, machine learning has become a major technology for the identification of models describing the behavior of molecules and physical phenomena. ML has been applied for the discovery of new materials and the prediction of new properties through quantum and statistical calculations, which has yielded many remarkable results. In addition, it is also used to solve problems related to materials science that require large calculation times and experiments that can be expensive and difficult to perform [2]

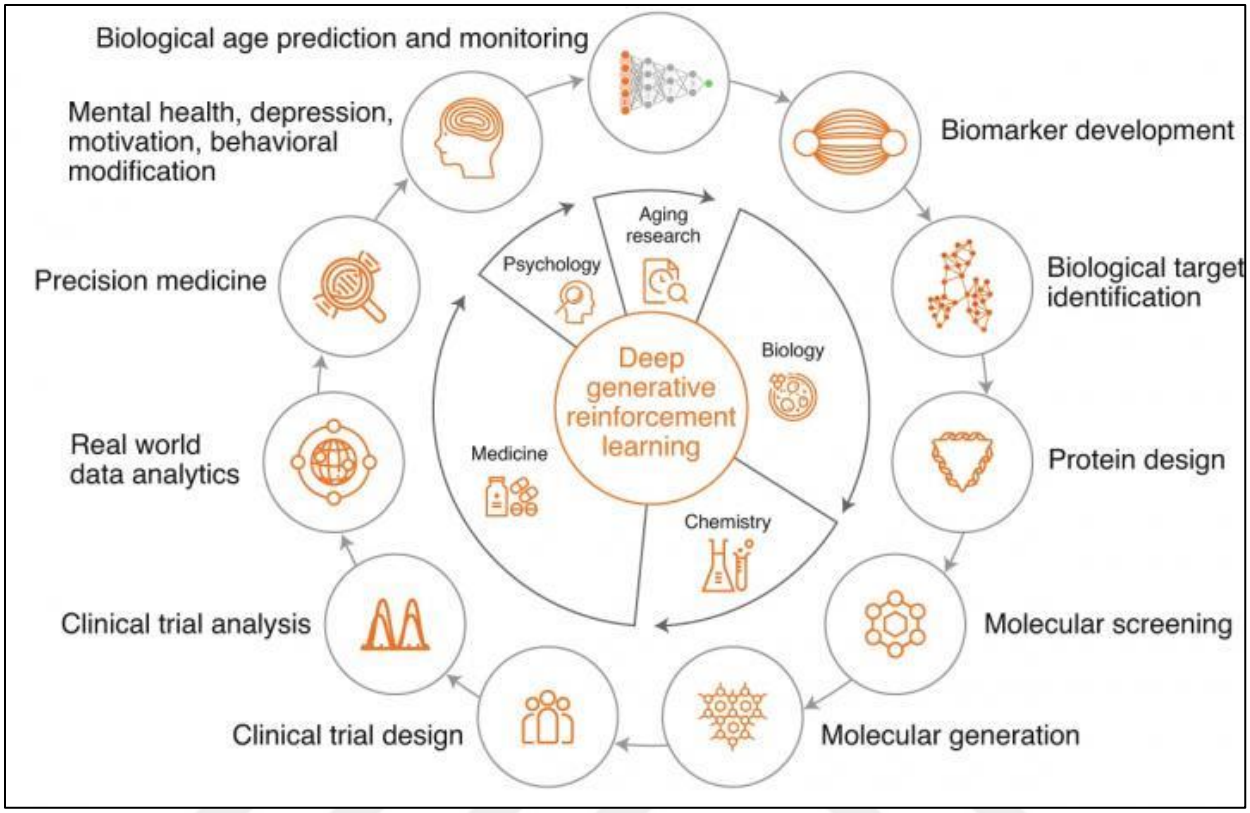

Figure 1.2: Machine learning and fields of application [2].

Twitter and LinkedIn also make use of a face recognition system that is founded on the concept of machine learning. Cameras used for surveillance also make use of this technology. When you conduct a search, algorithms on the server level will monitor how you respond to the results of the search. The results that search engines like Google produce can likewise be improved with the use of this technology. If you remain on a particular web page for a significant amount of time, the search engine will assume that the results it returns are pertinent to the query [2].

The fundamental idea behind Machine Learning is that we should provide computers with input information for them to analyze as well as the desired end result (ML). This enables the machine to develop a model that is capable of making inferences about new information based on historical data using the values that we offer. This is made possible by the fact that we provide the machine with the historical data. To obtain a particular program that can be executed on a computer in order to solve a problem or accomplish a task that has been clearly described is the objective of an analyst. Users of Machine Learning who are classified as "Data Scientists" do not try to construct a program on their own but rather collect the input data and desired objective values. This is

because Data Scientists do not seek to create a program. After that, they provide the computer the instructions necessary to locate a program (or algorithm) that computes an output value for each new value that is given to an input. When people are unable to solve complex problems, machine learning can be used to solve those difficulties automatically. However, in order to develop a more accurate model, the Data Scientist needs to always be prepared and able to try out a number of alternative strategies before settling on one that meets their requirements. The term "machine learning" can be put to use in either of two different contexts. The first approach is known as supervised learning, and it makes use of algorithms that are founded on input data that has been labeled with the outputs that are sought. If we have a collection of objects, and each one of those objects has a target value associated with it, then we need to learn a model that is capable of predicting the proper target value of a new item that has not been labeled with a target value. The second sort of learning is known as unsupervised learning, and it differs from supervised learning in that the data being learned from are not labeled in any way, which means that they do not have a target value.

In this particular situation, the algorithm discovers on its own the points of commonality between the data that it has been given. After that, it organizes the data into groups of data that are as distinct from one another as is practically possible by increasing the level of their homogeneity; each group contains data with characteristics that are comparable to those of the other groups. When we select the proper characteristics based on a bigger quantity of available input data, we move closer to identifying a more correct category or grouping of the data [2]. This is because the more data that is used to inform our decision, the more accurate our categorization will be. Numerical techniques are utilized in the field of materials science in order to successfully manage the complexity of the issues that are present in all areas of technology. In order to fulfill the requirements of a wide range of applications, researchers in academia and industry are conducting a search for novel materials that are of an exceptional quality. In this context, machine learning has emerged as an important tool for the aim of identifying models that characterize the behavior of molecules and physical phenomena. This context places machine learning in the context of the following: ML has been utilized for the purpose of locating new types of materials and forecasting their properties using quantum and statistical computations, both of which have led to the production of a plethora of impressive outcomes. In addition, ML has been utilized for the purpose of predicting the properties of new materials using quantum and statistical computations. In addition to this, it is used to find

solutions to issues that come up in the field of materials science and call for lengthy periods of calculation as well as tests that can be difficult to carry out and expensive to do so as well. This type of problem-solving is made possible by the use of this technology.

1.2 PLANNING THE PROBLEM

Since its launch in 2000, the Human Survival Database (HSD) has become the reference provider of survival estimates for more than forty countries or regions. These national indicators are widely used by demographers, as well as insurance companies, to obtain information on future survival. While the literature devoted to the sophistication of stochastic models of survival is abundant over the last decades, few contributions have focused on the reliability of the data. used in historical and predictive survival analysis.

After [2] speculated that the sensitivity of the method used to calculate survival rates from fertility shocks may account for certain peculiar cohort effects in England, researchers began to take notice of discrepancies in the available survival data. An approach to adjusting English and Welsh survival statistics using quarterly birth rate data was developed, and this theory was proven, by [2]. They created the Convexity Adjustment Ratio, which was later adapted by who demonstrated the universality of these anomalies using HSD data. Linking with the Human Fertility Database [2], the fertility counterpart of the HSD, allows for the systematic correction of these abnormalities, allowing for the construction of survival forecast tables.

There was a rise in the awareness of anomalies in survival data after the conjecture of [2], which suggested that particular cohort effects in England could be due to the sensitivity of the method of calculating survival rates from fertility shocks. This conjecture suggested that certain cohort effects in England could be due to the sensitivity of the method. This realization was brought about as a direct consequence of the hypothesis. In [2], research developed a method to update survival data for England and Wales based on quarterly birth rate data, provided evidence that supported this conjecture by demonstrating that the strategy is effective also provided evidence that supported this conjecture by demonstrating that the strategy is effective. The Convexity Adjustment Ratio, which was initially established by them, was later changed by [2], who focused on HSD data and demonstrated that these irregularities are present everywhere. [2] modified the Convexity Adjustment Ratio in this way. The construction of a link with the Human Fertility Database, which

is analogous to the HSD in terms of fertility has made it feasible to correct these defects in a methodical manner, making it possible to do so thanks to the fact that it has made it possible to do so. The creation of updated life tables will be made possible as a result of this.

The most current update to the HSD methods methodology [3] contains a change to the procedure that is used to generate survival rates in the database. This update was released in April of this year. Because of this shift, it is now possible to do a survival rate analysis that is more precise by utilizing monthly fertility data. However, the implementation of these strategies for correcting survival data is not possible in nations whose birth statistics are not available monthly.

The latest version of the HSD methods protocol introduces a change in the way survival rates are constructed in the database, improving their estimation by using monthly fertility data. However, these methods of correcting survival data do not apply to countries for which birth data by month are not available.

1.3 THESIS JUSTIFICATION

Survival forecasts are used in a wide variety of fields: for health policymaking, to guide pharmaceutical research, social security, retirement fund planning and life insurance, to name a few. Over the twentieth century, it is now well known that human survival has declined globally: in most industrialized countries, survival among adults and the elderly reveals a decreasing annual probability of death.

Moreover, in the literature on survival modeling, machine learning approaches are not very present. Among these approaches, [3] extended the Lee-Carter model to several populations using Deep Learning

The development of this technology for usage in countries for which we do not have access to the appropriate data is of particular importance to us, and we are particularly interested in seeing how it may be implemented there. Survival projections are utilized in a wide variety of settings, some of which include, but are not limited to, the formulation of public health policy, the direction of pharmacological research, the planning of social security and retirement funds, and the provision of life insurance. Other settings in which survival projections are utilized include the provision of life insurance. It is common knowledge at this point that the human survival rate has been on the

decline on a worldwide basis for the entirety of the twentieth century. Specifically, survival rates among adults and the elderly demonstrate a decreasing annual risk of dying in the majority of industrialized countries. This trend may be seen in survival rates. It is possible to observe this pattern in the survival rates.

In addition, there have not been many instances of machine learning approaches being incorporated into the research that has been done on survival modeling. This is something that has been lacking. To extend the Lee-Carter model to various populations, one of the techniques that was taken was to employ deep learning as one of the methods. This was one of the approaches that was taken.

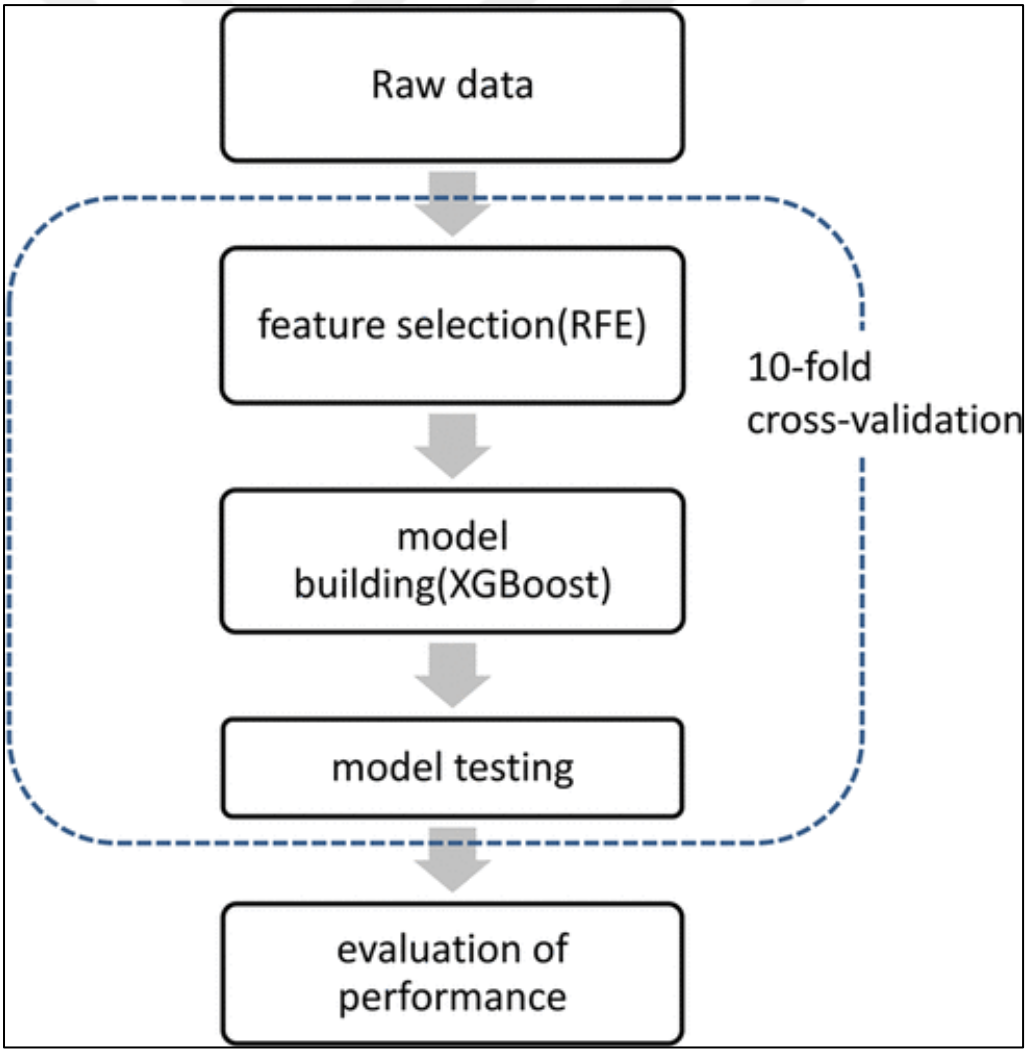

Figure 1.3: Abstract view of the proposed model [3].

1.4 OBJECTIVES

develop a model employing regression techniques and deep learning to predict the likelihood of survival for patients hospitalized with illnesses typically observed in intensive care units (ICUs). Survival models are a sort of statistical technique that is commonly employed in clinical research. The objective of these models is to identify potential risk variables and forecast the risk of various clinical outcomes, such as the overall survival of patients with a variety of diseases

1.5 THESIS ORGANIZATION

This thesis is divided into Five main chapters as follows:

In the first chapter, we introduce the most used algorithms in a Machine Learning process (statistical learning) whatever the problem posed.

In the second chapter we review some of the previous works in the state-of-the-art methods

In the third chapter we cite some databases that exist in the field of materials science, each of them contains specific properties. In addition, we mention the MATLAB language with which we will be able to extract the relevant data from our chosen databases, as well as the Scikit-Learn library that we will use in our machine learning process.

In the fourth chapter, three applications will be dedicated to machine learning. The first and the second part consist in creating a prediction model from a well-processed database.

Finally, the last part will be a statistical study on a dataset containing oxides, nitrites and copper-based materials.

2. LITERATURE REVIEW

2.1 CHAPTER OVERVIEW

This chapter addresses some of the concerns that have been raised regarding the overall survival rate. In the first part of this article, we will take a close look at the evolution of machine learning algorithms as well as their history.

In the next part, we will investigate the several ways in which one might pass away. The advanced machine learning subfield of artificial intelligence will be the topic of discussion for the course's concluding lecture.

The fourth half of this chapter is focused on exploring the numerous applications of regression as well as its ramifications. In the fifth segment, we will discuss the dangers that artificial intelligence poses. As a follow-up, in the following episode we will go over some basic principles for artificial intelligence.

2.2 RELATED WORKS

Because it is susceptible to structural factors such as socioeconomic advancement and fundamental living arrangements, the survival rate (MR) is an essential indication of mother and newborn health [3]. Between 1990 and 2017, the rate of infant mortality in Sub-Saharan African (SSA) countries dropped from 182 to 58 per 1000 live births [3]. Since the year 1990, we have witnessed a gradual decline that reached its nadir in the year 2017.

The primary objective of SDG-3 is to reduce the high rate of neonatal deaths to at least 12 per 1,000 live births and are under survival to at least 25 per 1,000 live births [3] [4]. This goal is part of the larger global effort to "ensure healthy lives and promote the well-being of all at all ages" by the year 2030. However, the rate is still deemed high.

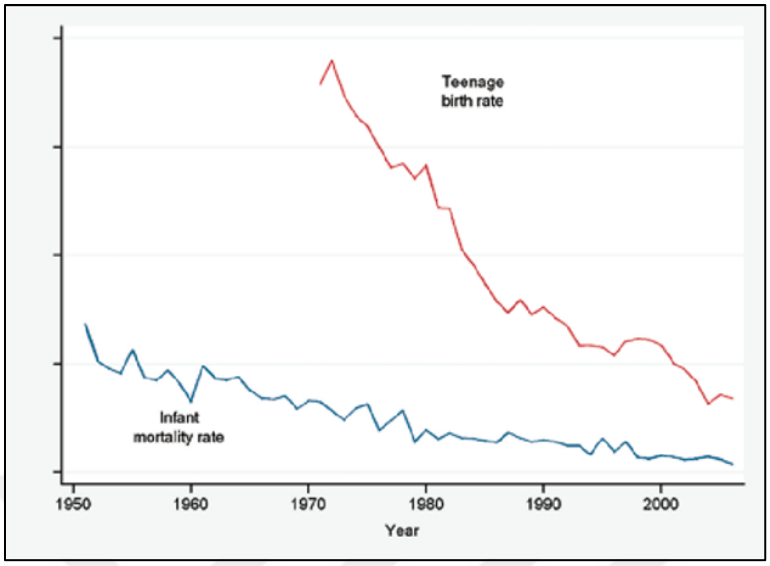

Figure 2.1: Predicting infant mortality with statistics [4].

The health of children has greatly improved as a direct result of the efforts made to reduce the prevalence of the risk factors that cause this health problem.

This has become a reality as a direct result of all of the laborious effort and meticulous preparation that went into it. Recent years have seen a surge in research that focuses on the potentially fruitful fields of artificial intelligence, machine learning (ML), and deep learning (DL) [3] [4]. It is feasible that this enhancement will become attainable if their plans are put into action. The amazing effectiveness of these technical advancements has been proven to by a wide variety of sectors, including medicine and public health [3][4].

Following the implementation of these strategies, the subsequent stage is to make use of AI software [5]. Previous research [6]–[11] that analyzed IM risk variables and their correlations to human survival utilized more typical statistical methods. Approaches like this include methods such as survival analysis, multivariate decomposition, and discrete-time logistic log-likelihood models, to name a few. When it comes to artificial intelligence, the performance of machine learning (ML) approaches is said to be superior to that of traditional ones when working with big datasets, according to a number of publications [3], [4].

The findings indicate that ML systems rapidly adjust to changing circumstances and become more effective. On the other hand, the overwhelming majority of these research have been carried out in countries that have a high standard of living [3], [4]. Researchers in sub-Saharan Africa (SSA)

concentrated on machine learning or traditional methods, but because of their small sample sizes, they were unable to provide an explanation for the vast range in infant survival rates found throughout the region [12]–[17]. This was due to the fact that researchers in SSA made use of antiquated methodologies and had insufficiently large sample sizes.

Discovering previously hidden patterns may be made significantly easier, more practical, and more essential with the use of machine learning approaches. These techniques depend on a priori assumptions made about the structure of the data, and as a result, they make use of processes that are not really implemented [18]–[20].

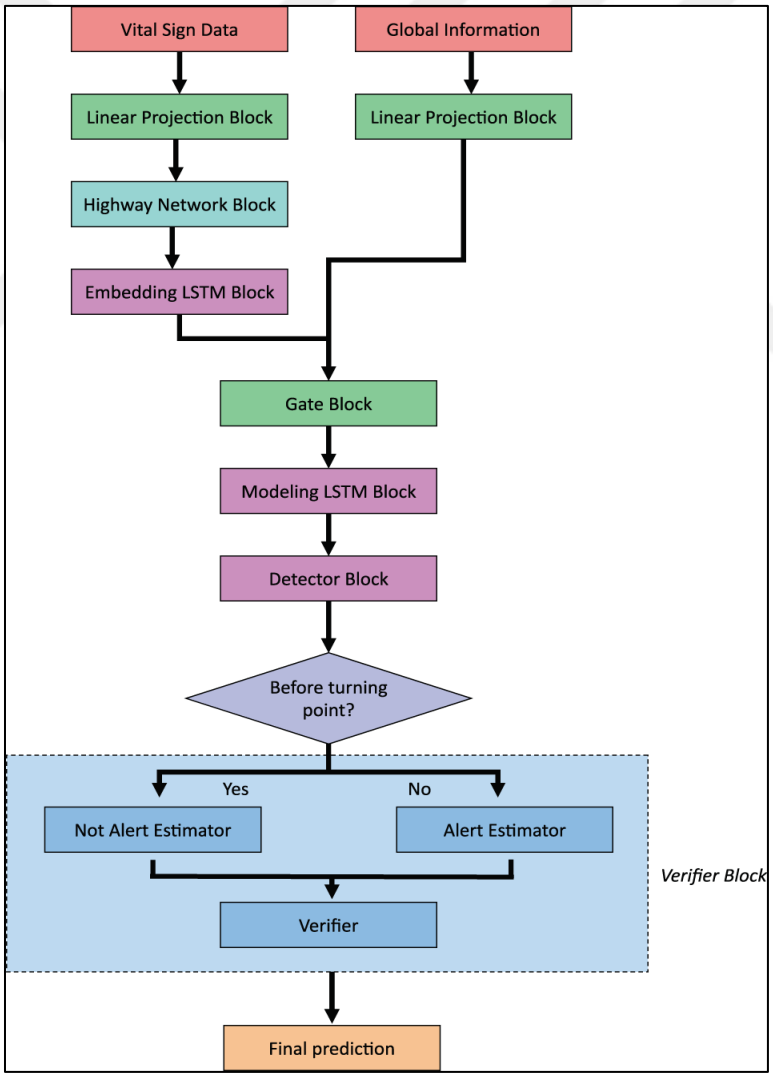

Figure 2.2: Process flow for predicting survival using machine learning [5].

They are an advantageous addition to the statistical tools that are already available. It is a highly helpful scientific approach for forecasting human mortality rates since Advanced Machine Learning (AML) can predict deaths that occur both during pregnancy and in neonates. It has also been demonstrated that AML may be used to forecast the likelihood of death in adults. The findings can be used by policymakers as a tool to influence the design of efficient health programs that would decrease exposure to possible threats [18]–[21].

Traditional analytic techniques, such as multiple logistic regression models, are unable to meet the better degree of prediction accuracy provided by AML systems when assessing the risk of human death [13–20]. Medical experts have looked at ML approaches for prognostication and survival prediction; however, ML methods have not yet been employed in community health settings, which is a wasted opportunity because this is where they can have the most impact [3, 22], [23]. In spite of the application of ML techniques.

The Logistic Regression model has an accuracy of 86.1 percent, while the K-nearest neighbors' model has an accuracy of 85.6 percent [24]–[26]. Random Forest could make accurate predictions between 67.2% and 97.1% of the time. The Random Forest model has an accuracy that can range anywhere from 67.2% to 97.1%. One of the many advantages that ML models have over more traditional statistical models is their flexibility in dealing with non-linear, complex data, as well as the various interactions that can occur between determinants and a contemporaneous series of events [2], [27]. This is just one of the many advantages that ML models have over more traditional statistical models. This is only one of several supporting reasons that may be made.

In addition to these advantages, machine learning models provide a variety of other advantages. The capability of machine learning to recognize when a woman or person is in a higher-risk scenario is one of the most important aspects that go into calculating the chance that a person will pass away. There are a lot of additional elements at play, but this is undeniably one of the more important ones. Machine learning algorithms may be used to unravel non-linear or binary relationships, which can be a challenge when trying to find correlations in large datasets. However, these algorithms are becoming increasingly sophisticated. When it is particularly difficult to identify connections, this proves to be quite beneficial. The aforementioned data sets are amenable to extremely accurate inferences and forecasts according to the aforementioned methodologies [3], [27]–[29].

Although making accurate estimates of human and baby survival may be challenging and often requires the cognitive ability of individuals, machine learning (ML) may be able to help with this. Because of its flexibility in accommodating a large range of differentiating characteristics, the linear discriminant analysis (LDA) model has been demonstrated to be the most effective model overall. This is the case because the LDA model is able to take these aspects into consideration. We utilized Decision Trees, Random Forests, Support Vector Machines, and Naive Bayes to distribute data on human fatalities and the underlying causes of those fatalities. In addition, using techniques from machine learning, they made predictions for the rates of newborn and child mortality for countries such as Ethiopia. The random forest outperforms other classifiers by a wide margin, with an accuracy of 98.43% on unbalanced train data and an average of 95.223% on balanced train data [29]. This is in comparison to other classifiers, which have an accuracy of just about 95%. The decision tree makes it abundantly evident that there is a significant connection between the rising population and the percentage of children who are surviving to adulthood. The fact that the constructed model has been used in a variety of contexts has contributed to its high level of respect [13]–[29].

When contrasted with the accuracy of other predictive classifiers, the Naive Bayes classifier's score of 98 percent [26]–[30] stood head and shoulders above the competition. This classifier came out on top with a ROC area of 0.92, making it the obvious victor. Nevertheless, evidence from [26]-[31] suggests that the decision tree model is the more reliable of the two. Using data mining techniques like neural networks and decision trees, researchers were able to calculate the likelihood of a child dying given certain environmental conditions like the number of windows in a house and the quantity of water present in the house. Researchers were able to calculate the likelihood of an infant dying using this approach. Preliminary research using machine learning (ML) approaches to predict neonatal and perinatal survival found that the algorithms employed to create these predictions enhanced prediction accuracy while significantly reduced bias [13]. According to these analyses, the algorithms employed to make these forecasts significantly enhanced prediction accuracy.

It would appear that research on machine learning (ML) has a significant impact, not only on the procedure for dealing with missing data, but also on the choice of variables. Because of this, certain persons were not included in the research project so that the findings would not be impacted in any

way by their absence. Although these predictive methods perform better than traditional models when used to large datasets [3], [4], there is a lack of research being done on them in SSA at the time [14]–[33]. Despite the fact that these methods have the ability to disclose unique correlations between components and differentiate composite connections, this is the case for the reasons stated above.

The national survival rate, often known as the MR, is an important indication of the health of both the mother and the child [4]. The phrase "maternal and infant survival rate" is what "MR" stands for in its abbreviated form. The MR is vital not only for mothers, but also for their children because of its sensitivity to structural variables such as socioeconomic advancement and critical living arrangements. This makes the MR important for both mothers and their children. Because of this, the MR is extremely important for both mothers and their children. As a result, the MR is significant to the activities of both towns. The infant mortality rate (IMR) in Sub-Saharan African (SSA) nations decreased to 58 per 1,000 live births in 2017 [6], down from 182 per 1,000 in the previous year. This is a drop from the prior estimate, which was 182 deaths; the current number is lower than that. It is essential to keep in mind that you should be discussing the infant survival rate whenever you are talking about the IMR. This overall trend toward worsening may be noticed between the years 1990 and 2017, which is the important time period.

The third Sustainable Development Goal (SDG-3) of the world aims to "ensure healthy lives and promote the well-being of all at all ages" by the year 2030, with the primary goal of lowering the high rate of neonatal deaths to at least 12 per 1,000 live births and increasing under-five survival to at least 25 per 1,000 live births [3, 4]. However, these numbers remain relatively high. The measures and strategies used to reduce the underlying causes of this health problem have had a profoundly positive effect on children's health. It is important to emphasize that this general trend toward better child health has been documented in a wide range of settings.

The primary goal of SDG-3 is to reduce the high rate of neonatal deaths to at least 12 per 1,000 live births and are under survival to at least 25 per 1,000 live births [3, 4]. This goal is part of the global effort to "ensure healthy lives and promote the well-being of all at all ages" by the year 2030. In an endeavor to lower the shockingly high newborn mortality rate, this objective is a key component. The rate is still deemed high, though. Efforts to address the underlying causes of this

health problem have led to major gains in children's health. Improvements in child health have been documented in a wide range of settings.

This end outcome is very much in large part due to the efforts that have been put in and the work that has been done up to this time. Recent research [3, 4] suggests that AI, ML, and DL may speed up this development. [Show citation] The many levels of education are collectively referred to as "tiers of learning." These three distinct modes of education are together referred to as "learning tiers," or simply "tiers" for short. It was shown that this possibility exists across all three of these distinct pedagogical approaches to the way education is delivered in schools. If and when they put their plans into action, they will improve their chances of reaching the milestones of achievement that they have set for themselves. The fields of medicine and public health are only two instances of businesses that have clearly established advantages from the technical breakthroughs that have been made [3, 6].

Following the execution of these methodological procedures, the subsequent step is to make use of computational intelligence [7]. This stage comes after the implementation of the numerous approaches that have been used up until this point. Studies of risk factors for infectious diseases (IM) and those connected with human survival [7]-[11] have typically been carried out using conventional methods. The survival analysis, the multivariate decomposition, and the discrete-time logistic log-likelihood models are some examples of the types of methodologies that may be categorized under this overarching category. When it comes to dealing with huge sample numbers, a number of AI specialists [3, 4] are of the opinion that machine learning (ML) approaches are preferable to traditional methods (AI).

Machine learning Techniques have the potential to rapidly adapt to new conditions and become more effective based on their performance. However, the great majority of these research have been conducted in nations that have a high standard of living [3, 4, 12]. Because of their small sample sizes, researchers in sub-Saharan Africa (SSA) concentrated on either machine learning or traditional methods; nevertheless, they were unable to provide an explanation for the vast range in infant survival rates that existed throughout the region [14–17]. The researchers relied on antiquated methodology and had too little of a sample size to arrive at any definitive findings from their work. This was due to the fact that the researchers, just like everyone else, relied on tried-and-true research methodologies and used an insufficiently large sample size. This was due to the

fact that the investigations conducted by the researchers relied largely on either traditional procedures or ML approaches. In this particular instance, the conventional method of educational practice. This was due to the fact that the researchers from the SSA relied on outdated methodologies and employed an insufficiently large sample size. Due to the low total number of participants in the sample, there was insufficient statistical power. They did not collect a sufficient number of samples, which is one of the factors that contributed to the development of this issue.

Discovering previously unknown patterns is made easier with the help of machine learning techniques, which offer novel, immediately useful, and highly efficient methods for making these discoveries. "Deep learning" is the umbrella term that has been given to the underlying mechanism that is shared by all of these different methods. The utilization of unknown processes and the formation of specific a priori assumptions regarding the character of the data serve as the foundation for these techniques [19], [20].

These recently proposed statistical procedures are supposed to work in conjunction with the tried-and-true statistical approaches that came before them in the line of statistical thought. It would appear that the branch of study known as Advanced Machine Learning (AML) is quite helpful when attempting to estimate death rates. This is as a result of the fact that AML can accurately forecast the death rates of both mothers and infants. Because AML may cause death in neonates, it is a very dangerous condition, which is why it is treated with such seriousness. In addition, AML is able to make an accurate prediction of the total number of fatalities that will take place. The findings have apparent repercussions for policymakers, who might make use of this information to design health measures that minimize risk factors for a number of illnesses and ailments [18]–[21].

It would be absurd to attempt to make a fair comparison between the precision with which AML systems predict the likelihood of human mortality and that with which conventional analytic techniques, such as multiple logistic regression models, predict the likelihood of human mortality. A fair comparison between the two would be absurd. This is because AML systems give an extremely high degree of precision, which is the reason why this is the case. The prognostication and survival prediction capabilities of ML approaches have been investigated by professionals in the medical field; however, ML methods have not yet been implemented in community health settings, where they may be a game-changing instrument [12]. This is still the case in spite of the

fact that professionals in the medical field have examined machine learning approaches. In spite of the fact that experts in the medical field have looked into solutions based on machine learning, this is still the case. This has taken place in spite of the extensive implementation of machine learning techniques.

Logistic Regression, on the other hand, has been calculated to have an accuracy of 86.1%, while the K-nearest Neighbors model has an accuracy of 85.6% [13]. Depending on the specifics of the problem at hand, the Random Forest model can achieve an accuracy of between 67.2% and 97.1%. Depending on the specifics of the event being modeled, the Random Forest model can have an accuracy ranging from 67.2 to 97.1 percent of the time. The ability of ML models to deal with non-linear, complex data, as well as many interactions between determinants and a contemporaneous series of events, is a major reason why they are preferred over traditional statistical models [14]. This is one of the main reasons why ML models are preferred over traditional statistical models. This feature is only one of the many benefits that machine learning models provide in comparison to more traditional statistical ones. Although this is true, it is not the only reason that ML models are favored over more conventional statistical ones. In fact, this is a far cry from the only reason. The conventional approaches to statistics come with a few prerequisites. ML models are favored over more traditional statistical models for a variety of reasons, and this is only one of those reasons. There is a possibility that more explanations exist for this phenomenon; the one that has been provided here is only one example.

Clients that purchase ML models are also eligible for a number of extra benefits and advantages. One of the most important aspects to consider when estimating the likelihood of a person passing away is the degree to which machine learning can differentiate between instances in which women and persons are in settings associated with increased danger. When attempting to estimate how long someone has left to live, this is one of the most important factors to consider. This is one of the key considerations that has been given attention. In spite of the fact that quite a few other factors might have an impact on this probability, it is still essential to keep in mind. In huge datasets, where finding correlations is notoriously difficult, approaches from the field of machine learning may be utilized to differentiate between linear and non-linear or binary connections. This comes in especially handy in situations in which it is difficult to locate any kind of relationship. This is especially helpful in situations in which determining correlations between the variables

being studied is difficult. When trying to discover the links between probable causes and the various effects those causes may have, this is an exceptionally valuable piece of information to have. This is especially helpful in situations in which it may be challenging to discover links between the variables that are being researched. When applied to the aforementioned data sets, these approaches are capable of producing estimates and projections that are quite precise [14]. Despite the fact that this task frequently requires the cognitive abilities of individuals, machine learning (ML) may be able to lend a hand in the search for reliable projections of human and infant survival. despite the fact that in most cases, the involvement of human intellect is required for ML. This is still the case in spite of the fact that ML might assist in the process of finding a solution. The linear discriminant analysis (LDA) model has been demonstrated to be the most effective model overall. This is likely due to the fact that the model is able to take into consideration a large variety of distinguishing characteristics. Because of this, the LDA model has been shown to be the most successful one overall. Moreover, this is one of the reasons why. It has been demonstrated to be the model with the highest rate of success; hence, it follows that this must be the situation. It has been found that the LDA model is the most successful of all models; hence, the success of this achievement may be attributed, at least in part, to the fact that this determination. The methods Decision Tree, Random Forest, Support Vector Machine, and Naive Bayes were utilized in order to successfully complete the task of data distribution on human mortality rates and the underlying causes. The data that was given emphasized the factors that have a role in the occurrence of such deaths. The transmission of this knowledge will hopefully assist in reducing the number of tragic events such as this one. Researchers were successful in accurately predicting the rates of infant and childhood mortality in countries like Ethiopia thanks to the application of machine learning technology, which were utilized to make the forecast. In addition to this, they offered their forecasts for the future rates of mortality. When applied to imbalanced train data, the random forest achieves an accuracy of 98.43%, whereas when applied to balanced train data, it achieves an accuracy of 95.22% [16]. This is a significant step forward in terms of efficiency when compared to the efficacy of other classifiers. This is a significant departure from the several other types. The decision tree demonstrates that there is a significant correlation between the growth of populations and the number of children who reach maturity each year. This conclusion was brought about as a result of contrasting the two rates. The fact that these two variables have a strong association is proof of this. The presence of a positive correlation between two variables is helpful in highlighting

the possible link that exists between them. When compared to earlier models, the one that was just created has an acceptance level that is unparalleled [16]. As a consequence of this, a very high level of reverence is accorded to it.

The Naive Bayes classifier came out on top after competing against a number of other types of predictive classifiers in a tournament. The fact that it has an accuracy rate of 98 percent [17] puts it in a league of its own when compared to other classifiers. The ROC area for this classifier was estimated to be 0.92, demonstrating beyond a reasonable doubt that it was the choice that was most suitable. In spite of this, the decision tree model has been shown to be capable of producing outcomes that are more accurate [18]. The researchers calculated the likelihood of a kid passing away based on environmental parameters like the number of windows in a home and the amount of water that was present in the dwelling by employing data mining techniques such as neural networks and decision trees. The researchers were able to calculate the chance of a youngster passing away by using environmental parameters such as the number of windows and the amount of water in the residence. In other words, the amount of water present in the house as well as the number of windows were taken into consideration. The utilization of environmental indicators helped to identify, to a certain extent, whether or not there was a threat present. The researchers were able to ascertain the probability of a child passing away as a direct consequence of their findings by doing mathematical computations. The algorithms used to generate these predictions improved prediction accuracy and decreased bias, according to preliminary studies that used machine learning (ML) methodologies to predict neonatal and perinatal survival [19]. These studies indicated that the algorithms used to generate these predictions improved prediction accuracy. According to the findings of the research, the use of prediction algorithms led to an improvement in the accuracy of forecasts. Research has shown that improving the algorithms that are used to make these predictions can enhance the accuracy of the forecasts that are generated using these algorithms. The findings of this research indicate that the utilization of algorithms led to an increase in the degree of accuracy displayed by the predictions made.

It is abundantly clear that research on machine learning (ML) has a significant impact on not only the strategy that is applied to missing data, but also the variables that are selected to fill in the gaps. This is how things work because of the way that ML functions. That this is the case makes perfect sense when one considers that ML is shorthand for machine learning. People were omitted from

the study as a direct result of this to prevent data-related bias, which is precisely why they were initially excluded, which is precisely why they were initially excluded, which is precisely why they were initially excluded, which is precisely why they were initially excluded, which is precisely why they were initially excluded, which is precisely why they were initially excluded. In spite of the fact that advanced prediction approaches perform better than standard models on big datasets [21], AML research is still extremely uncommon in SSA [22]. Even if these methods have the capacity to discern composite connections and find correlations between components that had not been discovered before, the situation is still the same as it was described earlier.

3. METHODOLOGY

3.1 CHAPTER INTRODUCTION

It is a well-known problem that has to be dealt with, and that problem is the incidence of Accidental brain injuries (TBI). The incidence of Accidental brain injury in Taiwan was 220.6 per 100,000 person-years in 2007-2008, whereas the in-hospital death rate was 10.7 per 100,000 person-years over the same time. The rate has been steadily rising, having gone from 521 per 100,000 person-years in 2001 to 824 per 100,000 person-years in 2010 [5]. Patients who have suffered an Accidental brain injury can have their prognosis predicted using a number of different variables. These variables include age [3], [8], gender [5,6], obesity [5], the Taiwan Triage and Acuity Scales (TTAS)[9], [2] , the Glasgow Coma Scales (GCS) [3], pupillary light reflex [10], and the degree of midline shift on a computed tomography (CT) scan [7]. The prognostic assessments that were developed by Steyerberg and colleagues [15] are based on the parameters that were considered at admission. These evaluations take into account a variety of factors, including the results of CT scans and laboratory data. In the case of patients with Accidental brain injury who are currently being evaluated at emergency triage stations, a workable model for predicting death has not yet been developed. This model would take into account a number of risk factors, but it would not require imaging scans or testing on blood sample specimens. As a consequence of this, the authors of this study are of the opinion that an attempt need to be made to enhance the standard of clinical treatment by conceiving of a trustworthy method for calculating the probability of passing away. In addition, professionals who specialize in critical care may use this information to make treatment decisions, such as deciding whether the patient should get conservative or aggressive therapy, and to educate the patient's family members.

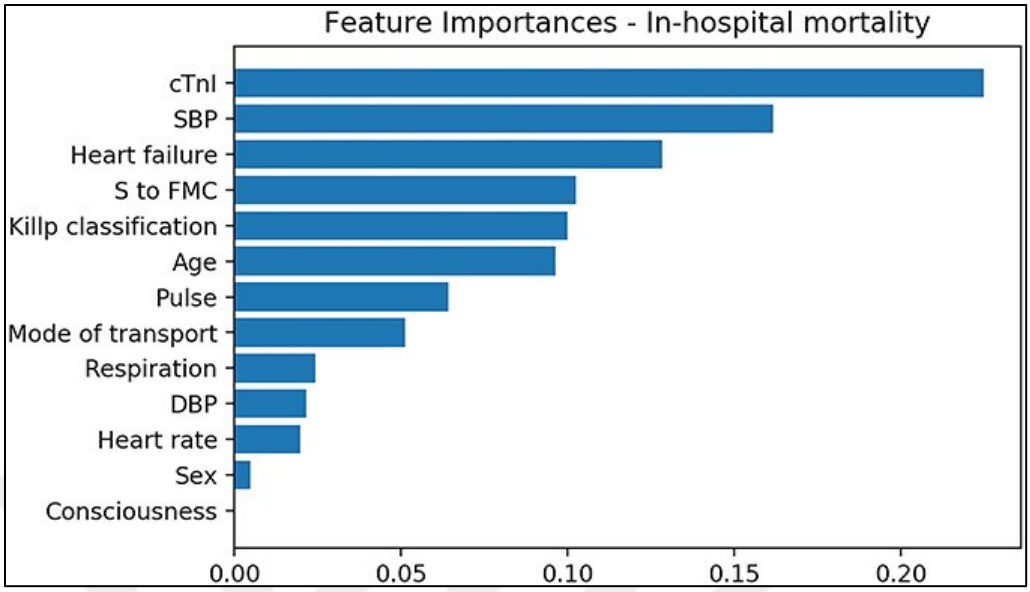

Figure 3.1: Feature based survival rate prediction

In clinical practice over the course of the past few years, a variety of statistical and artificial intelligence models have been put to use for the purposes of providing diagnostic suggestions [22], identifying adverse events [23] and surgical complications [24], and determining in-hospital survival. This particular collection of models contains a variety of models, some of which are referred to as logistic regression (LR), random forest (RF), support vector machines (SVM) [18], XGBoost [20], and multilayer perceptron (MLP) [21–27]. [16–17] Because modern high-performance computer systems now incorporate machine learning in addition to other core statistical theories, it is now possible to find patterns that may be utilized for identification and prediction. This is made possible as a result of recent technological advancements. This paves the path for new lines of research, which in turn paves the way for new discoveries. Because of this, it is now possible to identify patterns that are able to serve either one of these goals or both of these purposes simultaneously. The evidence that was provided in [26–29] suggests that the conditions had a considerable role in deciding how accurate their guesses were, and that this role was determined in large part by the conditions themselves. Inferences to this effect can be drawn from the fact that the criteria were brought up in conversation [26–29].

The issue is commonly overlooked during the triage process in emergency departments since there has been so little research done on the application of machine learning models to predict the survival of patients with accident brain injuries [30–34]. This is because there hasn't been a lot of

research done on the subject, which is why there isn't enough information. This is as a result of the fact that there has been an extremely limited amount of research carried out on the topic. Patients in the emergency department (ED) of the Chi Mei Center who were experiencing chest discomfort and major adverse cardiac events were the focus of the Chi Mei Center administration's efforts to develop a real-time prediction system, which was then used to target these patients [32]–[35]. The administration's efforts were directed toward developing a real-time prediction system that was then used to target these patients. This system was developed with the assistance of a methodology that is driven by large amounts of data in addition to machine learning, both of which were included in the hospital information system (HIS) of the center when it was being developed. Additionally, this system was built with the assistance of a methodology that is driven by large amounts of data. In addition to that, this system was developed with the assistance of a methodology that is based on a substantial quantity of data [36]. As a consequence of this fact, the construction of this system was feasible. It is now possible to use TTAS to appropriately prioritize treatment, which may help reduce the amount of time that is lost as a result of over-triage and enable emergency departments to make better use of the resources they have available to them [9].

3.2 MACHINE LEARNING ALGORITHMS

3.2.1 Supervised Learning

Supervised learning consists in using a set of data to predict statistically probable future events, which is to say, it from a prediction model from the events already previously predicted. ML models can be used in “prediction” or “classification” applications. There are two types of supervised learning problems:

i. Classification: Classification methods apply when the set of result values is discrete. This amounts to assigning a class (also called tag or label) for each input value. Classification techniques can be based on probabilistic hypotheses (example, naive Bayesian) [37], notions of proximity (example, k nearest neighbors) [38] or searches in spaces of hypotheses (example, decision trees) [8]. Choosing the right technique is important; it is necessary to be able to choose the most suitable method which will be able to best separate the training data.
ii. Naive Bayesian: Naive Bayesian classification is based on the assumption that all features are conditionally independent of each other [39]. This method is based on Bayes' theorem which

calculates the probability of an event using prior knowledge of related conditions. This theorem was discovered by an English statistician, Thomas Bayes, in the 18th century but he never published his work. After his death, his notes were edited and published by mathematician Richard Price [2]. The theorem is given by the following formula in figure 3.2.

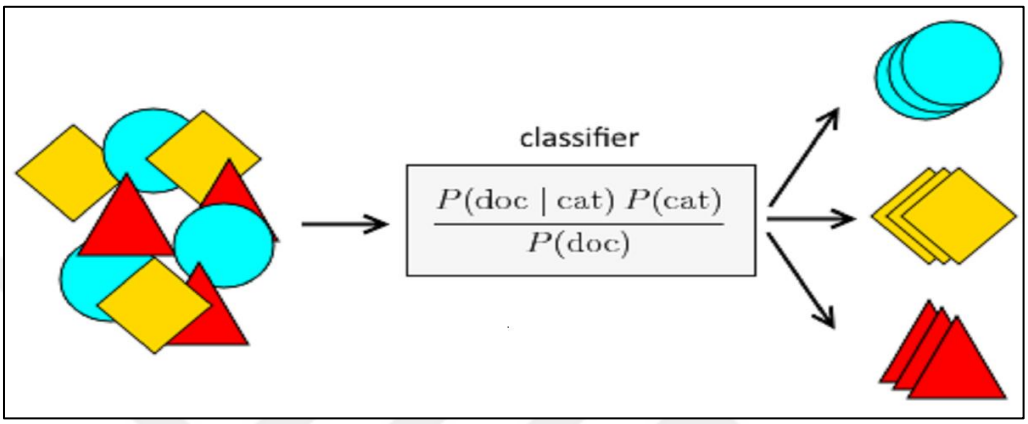

Figure 3.2: NB classification [40].

ii. k-nearest neighbors: Among the most basic machine learning algorithms is the k-nearest neighbor [40], often abbreviated as k-NN where k is a positive integer. In Data Science, this algorithm is widely used for data classification problems. But before embarking on this method, you should know that the calculations can be very expensive in terms of calculation time, so the data must be preprocessed. This method can also be used in regression problems. Consider for example the following classification problem: In the diagram below, there are green round objects and blue square objects. These belong to two different classes: the class of circles and the class of squares. When a new object is inserted into the space - in this case a red circle - we want the machine learning algorithm to classify the circle into a certain class [10]. If we choose k = 3, the algorithm seeks the three nearest neighbors of the red circle to be able to classify it either in the class of circles, or in the class of squares. In this case, the three nearest neighbors of the red circle are a square and two circles. Therefore, the algorithm will classify the sphere in the class of circles. In the k-NN method, the result is class membership. The algorithm stores all available cases and classifies any new object by checking its k-nearest neighbors. Then, the object is assigned to the class with which it has the most in common [10].

iii. Decision trees: Decision trees (DA) are a widely used classification tool. Its principle is based on the construction of a tree of limited size [8], [41]–[43]. Consider the simple example shown below. In this example, the vertex “Temperature > 37°C” represents the root. The leaves correspond to classes or decisions (also called terminal nodes), here “sick” or “healthy”. In addition, we call “Temperature > 37°C” and “Cough” the attributes or variables (also called intermediate nodes). ADs play a very important role in ML. They are able to handle both continuous and discrete variables and subsequently provide a clear indication for prediction or classification without performing much computation. ADs are so simple to understand and interpret, that they allow a better approximation regardless of the complexity of the data. The simplest tree is often the best, provided all other possible trees produce the same results. Their construction is carried out by dividing the tree from the top to the leaves (from top to bottom) by choosing at each stage a separation variable on a node, hence the segmentation criteria. For this, the most used algorithms are “the classification and regression tree” and “the iterative dichotomize” [12].

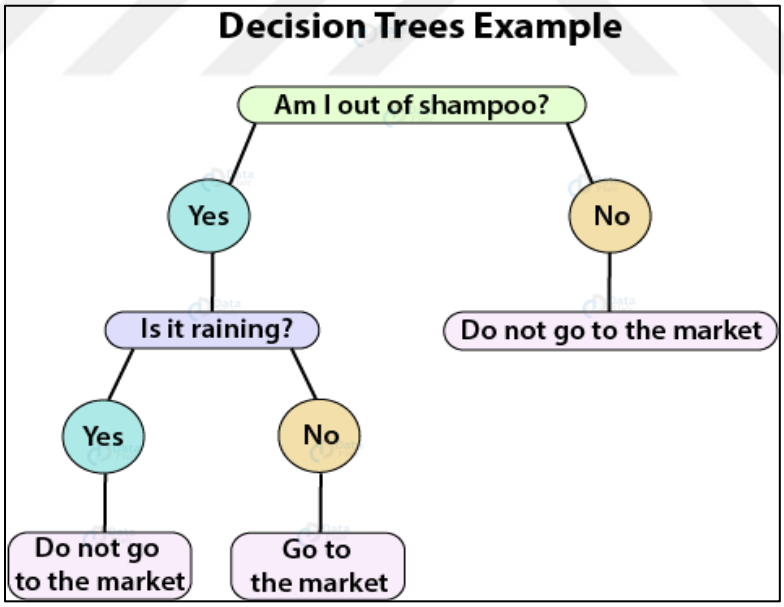

Figure 3.3: the iterative dichotomize classification.

iv. The classification and regression tree (CART): It was implemented by Leo Breiman in 1984 [3], [44]. The AD generated by CART is binary, i.e. the variable can only take two values (yes or no). CART's algorithm uses the Gini diversity index to assess splits in the data set. This index measures the frequency with which a random element of the set would be misclassified if its

label were chosen at random. The Gini index is given by the following equation, where c is the number of classes, p is the subset of node occurrences, and $P(k|p)$ is the probability of selecting an element of class c in the subset of node [12].

v. The Iterative Dichotomizer 3 (ID3): The algorithm was developed by Ross Quinlan and published in 1986. It involves Shannon's entropy which corresponds to the quantity of information delivered [45]. The objective is to build an AD from a set of data constituting the attributes. We must first have a root node, then we must determine the attribute that best classifies the training data, which corresponds to the highest information gain. We can calculate the degree of uncertainty using entropy. The latter is given by the following equation, where n is the number of possible outcomes and $P(xi)$ the probability of having position class i. Common values for b are 2, e, and 10, because the desired result must be positive [12].

vi. Artificial neural networks: The term neural network is a reference to neurobiology. Originally, this concept is inspired by the functioning of neurons in the human brain, essentially learning from experience [9]. Exactly how the human brain works is still a mystery, but some aspects are known. In particular, the fundamental element of the brain is a specific type of cell incapable of regenerating itself. These cells provide us with the ability to remember, think and react based on past experiences. The human body has 100 billion of these cells called neurons. Each of these neurons connects to 200,000 other neurons. Artificial neural networks attempt to reproduce only the most fundamental elements of this complex and powerful organism [12].

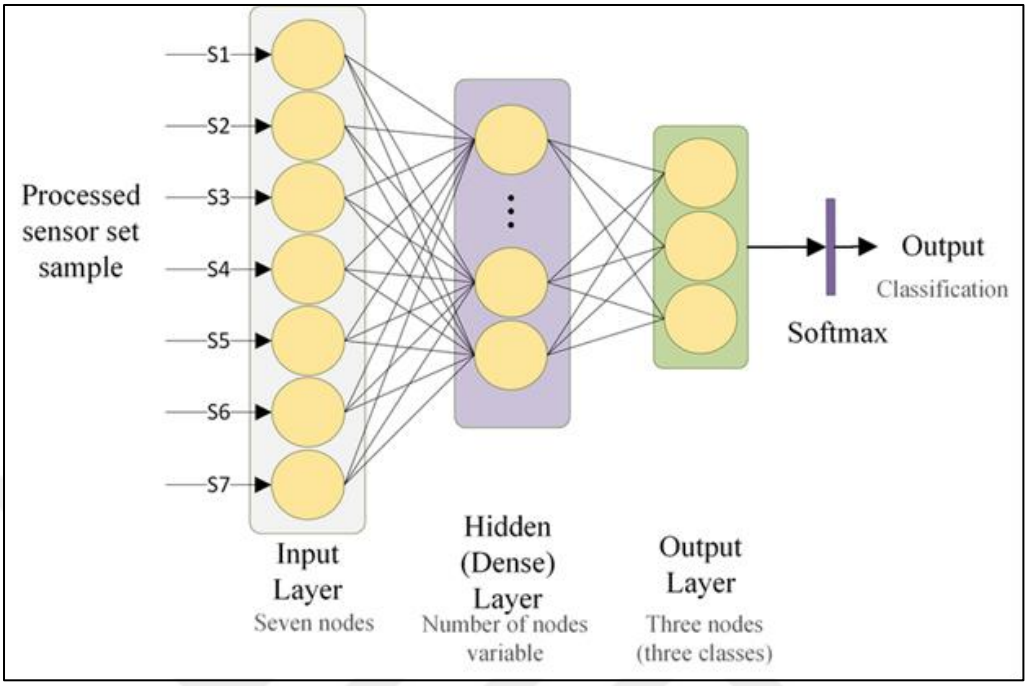

Figure 3.4: ANN structure [10].

A neural network is a hierarchical organization of interconnected neurons. The latter transmit a message or a signal to other neurons according to the input parameters received and form a complex network [10]. Above in figure 3.4 is a simplistic representation of a basic neural network: A neural network is generally composed of a succession of layers. This model has three sets of rules: multiplication, summation, and activation. Input data is consumed by the neurons of the first hidden layer. Each layer can have one or more neurons. The connection between two neurons of successive layers would have an associated weight which defines the influence of the input on the output of the following neuron and possibly on the overall final output. On each artificial neuron, each input value is multiplied by the corresponding weight [9] [10]. Then, the result obtained will be added to what is called the bias to apply to it at the end an activation function and the output value will therefore be of the form In a neural network, the initial weights would all be random during the formation of the model, the main unknown is its transfer function, it can be any mathematical function. The choice of this function depends on the problem posed. In the majority of cases, it is chosen nonlinear. This allows neural networks to learn nonlinear and complex transformations of data.

A neural network is a type of organizational structure that is characterized by the presence of neurons that are interconnected with one another in a number of distinct ways. These latter neurons establish a complex network with one another and communicate with one another by delivering a message or a signal to other neurons based on the input parameters they have received [10]. This communication takes place when one neuron delivers a message to another neuron based on the input parameters it has received. The transmission of a message or a signal is the means by which this communication is carried out. Figure 3.4, located at the top of the page, provides a representation of a fundamental neural network, albeit one that has been grossly oversimplified: The components that make up a neural network are almost always organized in the form of a hierarchical structure. This arrangement is the norm rather than the exception. Multiplication, totaling everything up, and activation are the three separate sets of rules that are incorporated into this paradigm. The neurons that are situated in the first hidden layer are the ones that are tasked with the responsibility of processing the input data. There is a wide variety of possibilities for the number of neurons that can be discovered in each layer, ranging from one to many. On each succeeding level, there would be a weight associated with each connection that was created between two neurons. These weights would increase as the number of levels increased. This weight would specify the influence that an input has on the output of the neuron that comes after it, as well as possibly on the output of the network as a whole. This influence would be on the output of the neuron that comes after it. This influence would be exerted upon the output of the neuron that is located immediately following it. The value of each input is multiplied by the weight that is linked with it, and this is done so in order to calculate the output of each artificial neuron [9], [10]. Next, the result that was achieved will be added to something that is known as the bias, and at the very end of the process, an activation function will be applied to it. This has a direct bearing on the shape that the final value will take, which is. The initial weights of a neural network are generated in a manner that is entirely arbitrary while a model is being developed using this type of computing system. One of the most important aspects of a neural network that is not fully understood is referred to as the transfer function, and this variable can take the form of any mathematical function. The nature of the issue that has to be solved should play a role in guiding your choice of function for this component. In the vast majority of instances, the nonlinear technique is selected as the method of operation that should be followed. As a direct consequence

of this, neural networks can now acquire the capability to learn nonlinear and sophisticated transformations of the data that they are given. This was not possible before.

3.2.2 Regression

Regression methods apply when the result that one seeks to estimate is a continuous value. In ML, regression is an important supervised learning tool for modeling and analyzing data. It is used in particular in statistics and economics [3].

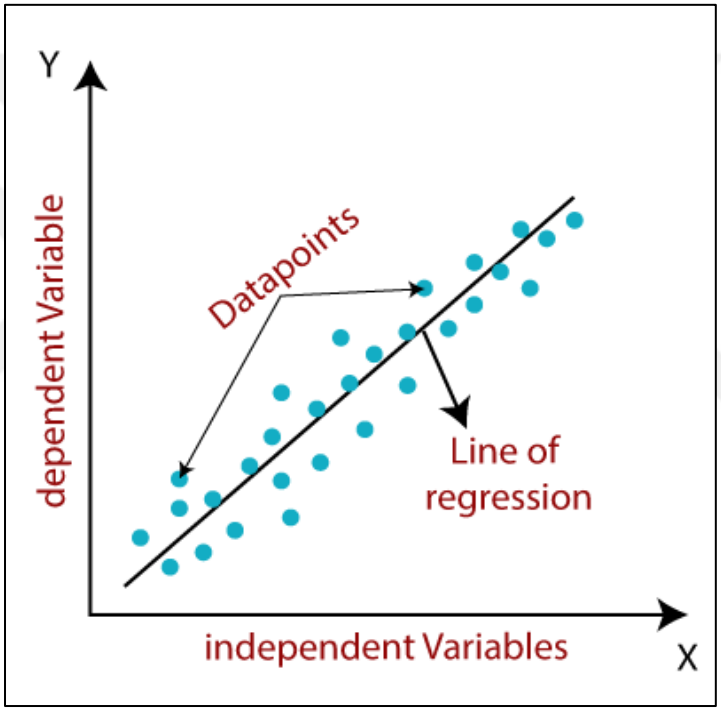

Figure 3.5: Regression in machine learning

i. Linear regression: A regression model is any model capable of establishing a linear relationship between a variable, called explained or dependent, and one or more variables, called explanatory or independent variables [46], [47]. The main goal is to fit a best line represented by a linear equation $Y = f(X) + \varepsilon$ in order to predict $\hat{Y} = \hat{f}(X)$ for any value of X. − Y represents the dependent variables. − X represents the independent variables. − ε represents the error or disturbance term. Linear regression is mainly divided into two categories: simple linear regression and multiple linear regression. Simple linear regression is characterized by an independent variable. On the other hand, multiple linear regression is characterized by at least two independent variables. In

statistics, the classic method is given by Ordinary Least Squares for solving simple regression problems; the best linear estimator according to the Gauss-Markov theorem, consists in calculating the sum of the squares of the residuals which we will denote by $scr\ (f)$ There is thus nonlinear regression, it is another form of analysis in which the data is modeled by a function of one or more independent variables, which is a nonlinear combination of the model parameters.

ii. Logistic regression: it was developed by statistician David Cox in 1958 [12]. Tea logistic model is a statistical model that uses a logistic function, it is also used in classification problems. Logistic regression does not require a linear relationship between dependent and independent variables, but it does require large sample sizes in order to have more precision when estimating likelihood. Logistic regression can be binary or multinomial. Binary logistic regression deals with situations in which the observed outcome for a dependent variable can only have two possible types, for example "sick" or "healthy", these two possibilities are labeled by "0" and "1". Multinomial logistic regression is for situations where the outcome may have three or more possible types that are unordered, e.g. "disease A" versus "disease B" versus "disease C". Logistic regression seeks to:

   i. Model the probability of an event occurring based on the values of the independent variables, which can be categorical or numerical.
   ii. estimate the probability that an event will occur for a randomly selected observation relative to the probability that the event will not occur.
   iii. predict the effect of a series of variables on a binary response variable.
   iv. classify observations by estimating the probability that an observation is in a particular category.

   The probability to be estimated during a logistic regression is represented by the inverse of the function “logit. ". The latter is equivalent to a linear function of the independent variables, given by the following formula

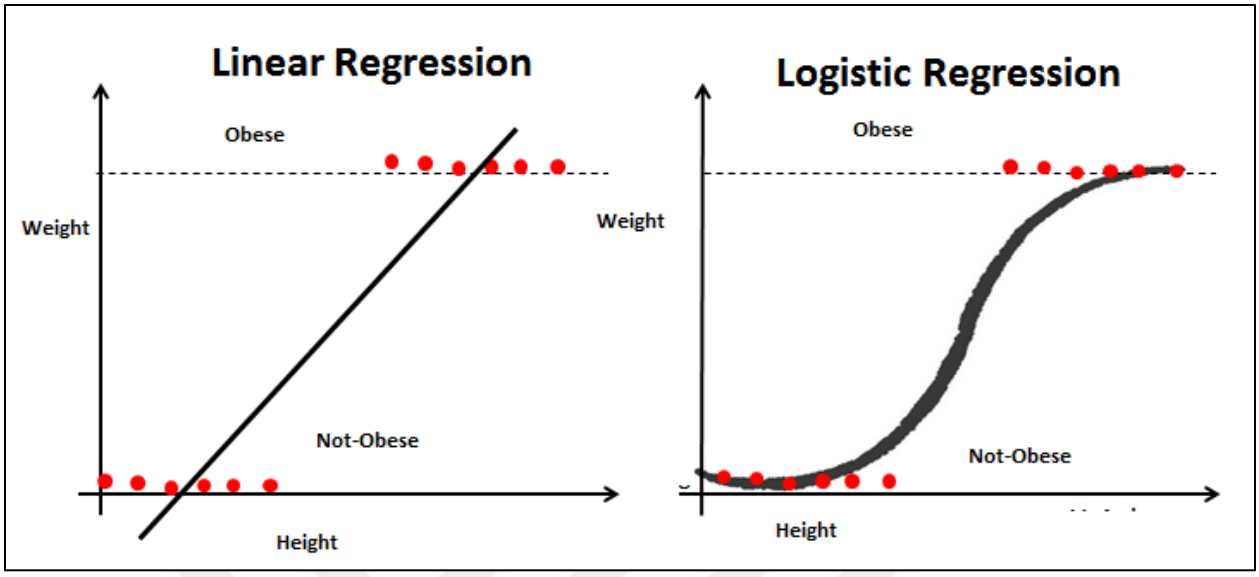

Figure 3.6: Logistic vs linear regression

Predicting Y' = f'(X) at any value of X requires finding the best line, which may be expressed as a linear equation of the form Y = f(X) + [11]. This is the primary objective, which can be attained by determining the line that will be represented by the linear equation. The dependent variables are denoted by Y in this equation. Error or the unforeseen event are both represented by the term "disturbance" in this equation, where X stands for the independent variable. Simple linear regression and multiple linear regression are the two main classifications utilized when discussing linear regression. Concurrent execution of both forms of linear regression is possible. In contrast to other statistical methods, simple linear regression is distinguished by its focus on the independent variable. In contrast, multiple linear regression can be distinguished from simple linear regression by the presence of more than one independent variable. In the realm of statistics, the Ordinary Least Squares technique is frequently considered to be the "gold standard" when it comes to addressing fundamental regression concerns. According to the Gauss-Markov theorem, the best linear estimator can be found by adding up the squares of the residuals (hence referred to as scr) (f). The best linear estimator can be obtained by adding the above values together. Accordingly, nonlinear regression exists as a supplementary method of analysis in which the data is represented by a function of one or more independent variables, where the function is a nonlinear combination of the model parameters. In nonlinear regression, the data is modeled as a function of several independent factors. Logistic regression, type B. In 1958 [12], statistician David Cox put forth the original idea of logistic regression. There are two distinct types of logistic regression:

The tea logistic model is a logistic function-based statistical model with applications in supply-chain management and product-classification challenges. While a linear relationship between dependent and independent variables is not necessary for logistic regression, larger sample sizes are preferred for more accurate probability predictions. When using logistic regression, the dependence and independence structure need not be linear. Binary or multinomial models can be used in logistic regression. Binary logistic regression is the suitable statistical procedure to use when the observed outcome for a dependent variable can only take on two distinct forms, such as "ill" or "healthy." These two possibilities are represented by the numbers 0 and 1. Multinomial logistic regression is useful when the outcome may be one of three or more unordered categories, such as "illness A" versus "sickness B" versus "disease C." Some of the things that logistic regression tries to do are: Assign categories to observations based on the likelihood that a certain event will or will not occur; Model the probability of an event occurring based on the values of the independent variables, which can be categorical or numerical in nature; Predict the effect of a set of variables on a response variable that can only take one of two values.

3.2.3 Unsupervised Learning

In unsupervised learning there are no output values, it is a matter of finding hidden structures from of a set of data that must be grouped, hence the term "clustering". The purpose of this type of learning is to separate the data into groups or categories.

i. Clustering: Clustering is an unsupervised machine learning technique used for clustering unlabeled data in many fields. If we have a finite number of data points and we want to classify them into groups so that each group contains data points with similar properties and/or characteristics. The main problem that arises in these algorithms is the choice of properties to take into account during clustering [13]. One of the most widely used clustering algorithms is “K-Means”.
ii. K-Means: This is the best-known classification algorithm. Its principle is simple, easy to understand and to implement in code. First, we select a certain number of groups and then, randomly, we initialize the center associated with each group. It is better to start by analyzing the data present globally and try to identify distinct groups in order to better determine the number of classes to use. Each point is classified by calculating the distance between this point

and the center of each group. Therefore, the point will be moved to the group whose center is closest. For each class, a new center is calculated as the average of all points in that group. After each iteration, the centers move slowly and the total distance between each point and the center assigned to it becomes smaller and smaller. The above steps will be repeated for a set number of iterations or until the cluster centers no longer change. K-Means ensures convergence towards a local optimum. However, this does not necessarily have to be the best global solution (global optimum), for this reason, one can initialize the cluster centers several times randomly, and then select the cycle that gives the best results [14]

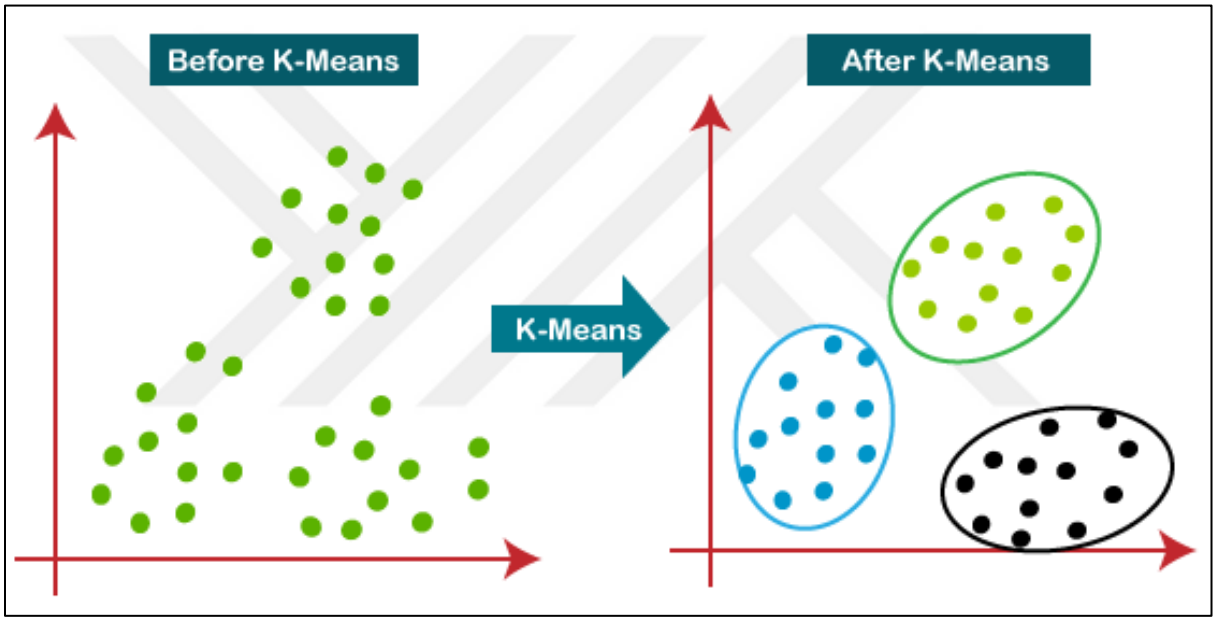

Figure 3.7: K-means clustering workflow.

iii. T -distributed stochastic neighbor embedding (T-SNE): It is an unsupervised linear technique, developed by Laurens Van der Maatens and Geoffrey Hinton in 2008. T-SNE is a dimension reduction method, it transforms the representation of multidimensional data in two or three dimensions and therefore gives an idea of how data is arranged in a high-dimensional space. T-SNE finds patterns from data by identifying groups (clusters) containing data that share similar characteristics. But it is not entirely a clustering algorithm, it is essentially a technique for exploring and visualizing high-dimensional data [15]. T-SNE converts high-dimensional Euclidean distances between data points into conditional probabilities representing similar features. The similarity of the data point xj with the data point xi is the conditional probability $pj|i$ that xi would choose xj as its neighbor if the neighbors were chosen with respect to their

probability density. For close data points, $pj|i$ is relatively large and vice versa [16]. Mathematically, the conditional probability $pj|i$ is given by Where σi is the variance of the Gaussian centered on the data point xi . For points obtained in the reduced-dimensional space yi and yj , it is possible to calculate a similar conditional probability, denoted $qj|i$. In T-SNE, a "t-Student" distribution (with one degree of freedom, identical to a Cauchy distribution) is used to measure the similarity between points in the reduced-dimensional space to allow dissimilar objects to be modeled [16]. More precisely $qj|i$ is defined by Logically, the conditional probabilities $pi|i$ and $qi|i$ must be equal for a perfect representation of the similarity of the data points in the different dimensional spaces, i.e. the difference between $pi|i$ and $qi|i$ must equal zero for perfect modeling in low-dimensional space. T-SNE attempts to minimize this conditional probability difference [16].

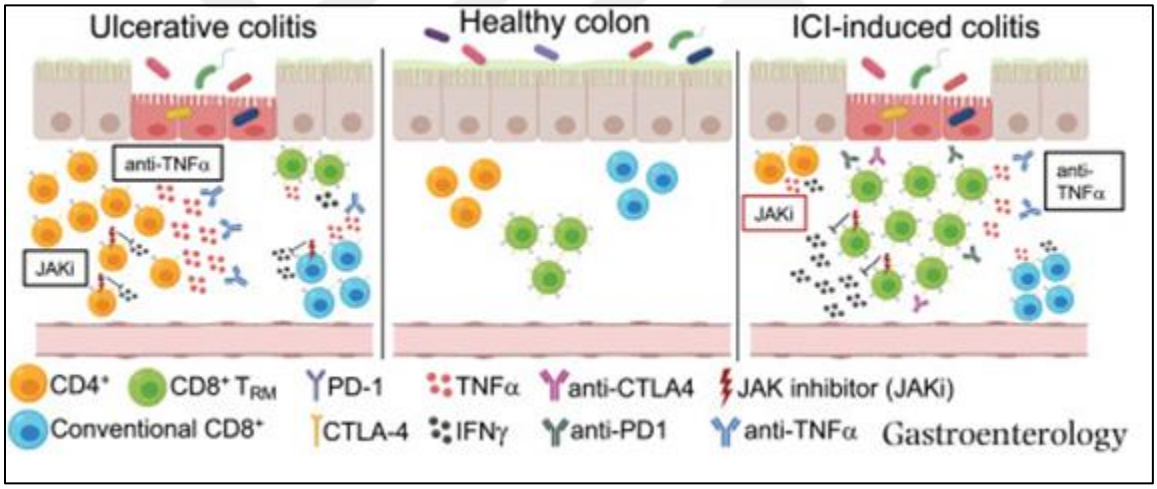

Figure 3.8: T -distributed stochastic neighbor embedding [16].

The locations of the yi points in the reduced-dimensional space are determined by minimizing the non-symmetric Kullback-Leibler divergence of the Q-distribution, represented by the following formula 3 Concept of distance: In order to measure the similarity between the elements of a data set, the choice of distance plays a very important role. It is necessary to identify, how the data are related to each other, and how the data are different or similar to each other [15]. For the calculation of distance in the K-Means method, several notions exist. However, Euclidean distance gives the best result. 3.1 Euclidean distance: Euclidean distance calculates the square root of the difference

between the coordinates of a pair of objects (points or classes). If we consider two points A and B, with respective coordinates (XA, YA) and (XB, YB) [15] [17].

3.3 METHODOLOGY AND LOGIC

A goal of care conversation (GCC) is a discussion between a patient (or a proxy) and his or her healthcare professionals to guarantee that the individual receives treatment in line with his or her preferences [4]. hospitalized people towards the end of life do not experience accidents as frequently as they should [5]. A STR helps to prevent overtreatment in this demographic by ensuring that in-hospital care is tailored to the individual's preferences and by disclosing a bad prognosis to the patient if he is unaware of it. The quality of care suffers, and the patient may receive more intensive treatment than they want if this conversation is delayed [11].

Limiting aspects of STRs include the perceived ambiguity of the prognosis by doctors and the time needed to adequately carry out such a talk [4] [9]. It is evident that prioritizing the discussion of care objectives requires an estimation of the prognosis. However, doctors often have unrealistic expectations about their patients' prognoses [2], which can lead to missed chances to help patients and their loved one’s plan for end-of-life care. Furthermore, initiatives at the state, provincial, and federal levels are under way to expand access to palliative care. For many terminally ill patients, the final days of treatment include an ever-increasing barrage of life-saving measures.

Both doctors and patients know that this course of treatment goes against the wishes of the vast majority of people who would rather die in a palliative care setting where the focus is on alleviating suffering. As a result, the first step in providing palliative care is to identify terminally ill individuals who might benefit from it. It is recommended that a STR be performed as soon as possible after this initial step in the hospital setting. In addition, if a patient is informed of his prognosis, it is important to determine whether or not his preferred level of care is compatible with a palliative care plan. Hospital staff might save time on two interconnected tasks—completing STRs and referring terminally ill patients to the right palliative care programs—if they knew which patients were nearing the end of their lives.

The patient (or the patient's proxy) and the treating healthcare professionals should have a conversation about the patient's desired outcomes. The GCC's major responsibility is to ensure that

the patient's treatment is always being carried out in line with the patient's stated wishes. A GCC is the same thing as a discussion that centers on the patient's desired outcomes from treatment. "GCC" is short for "a conversation about the patient's care goals," which is what the dialogue refers to. It's crucial to highlight the fact that hospitalized patients who are extremely near to the end of their life have a far lower accident risk than they should [5]. Such a discovery demands serious examination. A STR serves a dual purpose for these patients, who often get excessive treatment. To begin, it ensures that the patient's wishes are respected at all times during hospital stays. Second, it confirms what the patient may or may not have suspected all along: that his prognosis is not good. Overtreatment is more common among these patients than among others. To rephrase, a STR can help curb the overmedication of patients. The patient's overall treatment may suffer, and they may be forced to endure more therapy than they would want if you fail to have this dialogue with them at the proper moment [11].

These are only two of the numerous reasons why STRs don't work as well as they could [8]. The uncertainty of the prognosis, according to medical experts, and the length of time needed to have a fruitful conversation about such issues are also factors included in [9]. No one should be shocked to hear that a prognosis prediction is necessary for rational discussion of treatment objectives. However, medical professionals frequently exaggerate their patients' prognoses [2]. As a result, patients and their loved ones may miss out on chances to gain knowledge about the options available to them in terms of end-of-life care. In their optimistic assessments of their patients' prognoses, physicians often err [2]. As this is occurring, efforts are being undertaken at the local, state, and federal levels to streamline access to palliative care. It is standard practice to provide more of the treatments meant to prolong a terminally ill patient's life, and to keep doing so right up until the patient dies. In cases when it has been decided that the patient has a very limited time left to live, this therapeutic option may be made available.

Clinicians and patients alike are aware that this care plan is at odds with the wishes of the vast majority of patients, who would prefer to die in a palliative care setting where the focus is on comfort care rather than cure. When asked, most people say they'd rather die here than anywhere else. Most individuals in this world want to die right now and right now. This person has indicated a want to pass away in a respectful and respectable fashion. Therefore, the first step in providing palliative care is identifying patients whose lives are rapidly approaching their end and who might

benefit from it. Locating these people is a top priority. The hospital's next move, after making a diagnosis, should be to perform a STR as quickly as feasible. It is crucial to determine whether the patient's desired level of treatment is in line with the palliative care approach before informing the patient of his prognosis. After receiving a diagnosis, this needs to be done immediately. This issue must be fixed before a prognosis can be given for the patient. From this vantage point, it may be easier to do two related tasks if medical staff knew which patients were nearing the end of their lives. Patients must be referred to palliative care specialists and STRs must be filled out as part of these duties. The ability to identify which hospitalized patients are nearing the end of their life might greatly facilitate the measures.

3.4 SURVIVAL STATISTICS

The statistics presented in clinical research articles in surgery are evolving with the use of increasingly complex methods such as parametric and non-parametric tests, various types of regression analyses, and survival analyses [11]. While 35% of the articles published in 1985 in the Annals of Surgery and Archives of Surgery did not include any statistical analysis, we find less than 10% in 20032. A study evaluating the use of statistical methods in the New England Journal of Medicine reports that a reader with limited knowledge of a few descriptive statistics (percentage, mean, standard deviation) has full access to 27% of articles published in 1978-1978 against only 12% in 2004 Among the statistical methods used in 2004, 66% are survival analyses, followed by epidemiological statistics (relative risk, odds ratio, measures of association, sensitivity/specificity:

53%) and contingency tables (Chi- square, Fisher's exact test, McNemar's test; 43%) Paradoxically to this more frequent use of biostatistics, we note that 27% of a representative sample of 187 studies published in 2003 [3] in five major surgical journals incorrectly selected or reported their statistical methods: inappropriate t-test (confusion between t-test for independent vs paired data, or used for proportions when indicated for continuous variables); inappropriate Chi-square test (eg, used for small samples when Fisher's exact test should have been used); parametric test used for data that is not normally distributed when a non-parametric test should have been used and vice versa; inappropriate regression model (linear regression used for a binary dependent variable, instead of continuous, or logistic regression used for a continuous dependent variable, instead of binary);

tests that do not take into account the multiple comparisons made where it is theoretically possible to obtain 5% of the time, by chance, a "statistically significant" result for the sample of subjects studied when in reality there is no difference between the groups of subjects studied for the population to which we wish our inferences to apply.

Because researchers are using more complicated methods, such as parametric and non-parametric tests, various kinds of regression analyses, and survival analyses [3], it is becoming more difficult to comprehend the statistics that are presented in clinical research papers that pertain to the field of surgery. This is due to the fact that utilizing these methods enables the collection of a greater quantity of data. We discovered that fewer than ten percent of the publications that were published in the Annals of Surgery and the Archives of Surgery in 2003 did not contain any form of statistical analysis. This represents a significant improvement from 1985, when 35% of papers were issued without having undergone any kind of statistical analysis before being made public. Only 12% of the articles published in 2004 can be fully understood by a reader with only a basic understanding of descriptive statistics such as percentage, mean, and standard deviation, whereas 27% of the articles published in 1978-1979 in the New England Journal of Medicine can be fully understood by such a reader. This is because in 1978-1979, the New England Journal of Medicine encouraged its readers to send in their descriptive statistics in the form of percentages, averages, and standard deviations. The reason for this is that the New England Journal of Medicine asked its readers to send in their data. This information was gleaned via an analysis of the publication that investigated the statistical methodologies that were utilized. Survival analyses made up 66% of all statistical methods used in 2004, followed by epidemiological statistics (53%; relative risk, odds ratio, measures of association, sensitivity/specificity) and contingency tables (43%; Chi-square, Fisher's exact test, McNemar's test). In 2004, survival analyses were the most popular statistical method. 2004 was the year that saw the most utilization of the statistical procedure known as the survival analysis. We discovered that 27% of a representative sample of 187 papers published in 2003 [3] in five reputable surgical journals either utilized the incorrect statistical methods or offered incorrect information about them. This was the case for 27% of the publications in the sample. This goes against the current trend of more people using biostatistics. inappropriate t-test (such as mixing up the t-test for independent data and the t-test for paired data, or using it for proportions when it should have been used for continuous variables); inappropriate Chi-square test (such as using it for small samples when Fisher's exact test should have been used); parametric test used

for data that is not normally distributed when a non-parametric test should have been use. inappropriate test for continuous variables on data that is not normally distributed, a non-parametric test is utilized even if it would have been more appropriate to apply a parametric test. We utilized a regression model that wasn't correct (linear regression for a binary dependent variable when logistic regression would have been a preferable choice); we applied the incorrect regression model.

The descriptive statistical methods that were employed for continuous variables were found to be incorrect almost half of the time, according to the findings of another study [12] that looked at 144 publications that were published in 2005 by four of the most significant surgical journals. The findings of this study were derived from an examination of the aforementioned major surgical publications. It has been determined, following an investigation into the level of methodological rigor shown by the 653 articles that have been published in the eight surgical journals with the highest citation indexes, that the level of methodological rigor shown by the articles is not adequate.

Another analysis of 144 articles published in 2005 in four major surgical journals reports that the descriptive statistics methods used for continuous variables are inappropriate in 50% of case [12]. The methodological quality of 653 articles published in eight of the surgical journals with the best citation indexes reveals that the methodological quality of the articles is low (example: vague objectives, suboptimal or unspecified study design, absence of eligibility criteria and justification of the number of patients studied) [9].

3.5 ACCIDENT SURVIVAL RATE

According to the 2017 WHO road safety technical package [4,] there are over 1.2 million individuals who lose their lives as a consequence of traffic accidents every year all over the world, and millions more are injured as a result of these collisions. Accidents involving motor vehicles are currently, as reported by the World Health Organization, the ninth largest cause of mortality worldwide. According to figures compiled by the United Nations on the subject of the maturing of the world's population, in 2017 there were more than twice as many people aged 60 or older as there were in 1980. According to projections [5, the number of persons who are above the age of 60 will have more than doubled by the year 2050].

To be able to accommodate the inflow of additional inhabitants, the transportation infrastructure in many nations will need to be upgraded and improved. It's possible that as the population ages, there may be more old persons who require assistance going around. According to statistics from the National Highway Traffic Safety Administration for the year 2015, there has been a significant increase in the number of individuals aged 65 and older who are in possession of valid driver's licenses.

When there is an accident involving vehicles, drivers have a lower chance of being at fault, but they have a higher chance of being injured than their younger counterparts. According to recent data, there is an increase in the number of elderly drivers who are losing their lives in automobile accidents on EU roads. The elderly is at a greater risk of injury or death in accidents involving excessive speed as compared to accidents involving inexperienced drivers [3].

It is anticipated that in the not-too-distant future, the mitigation of risk factors related with traffic accidents caused by drivers would become an increasingly difficult task for transportation networks located all over the world. Finding out what factors might lead to an accident is the most important responsibility of a traffic safety specialist [12].

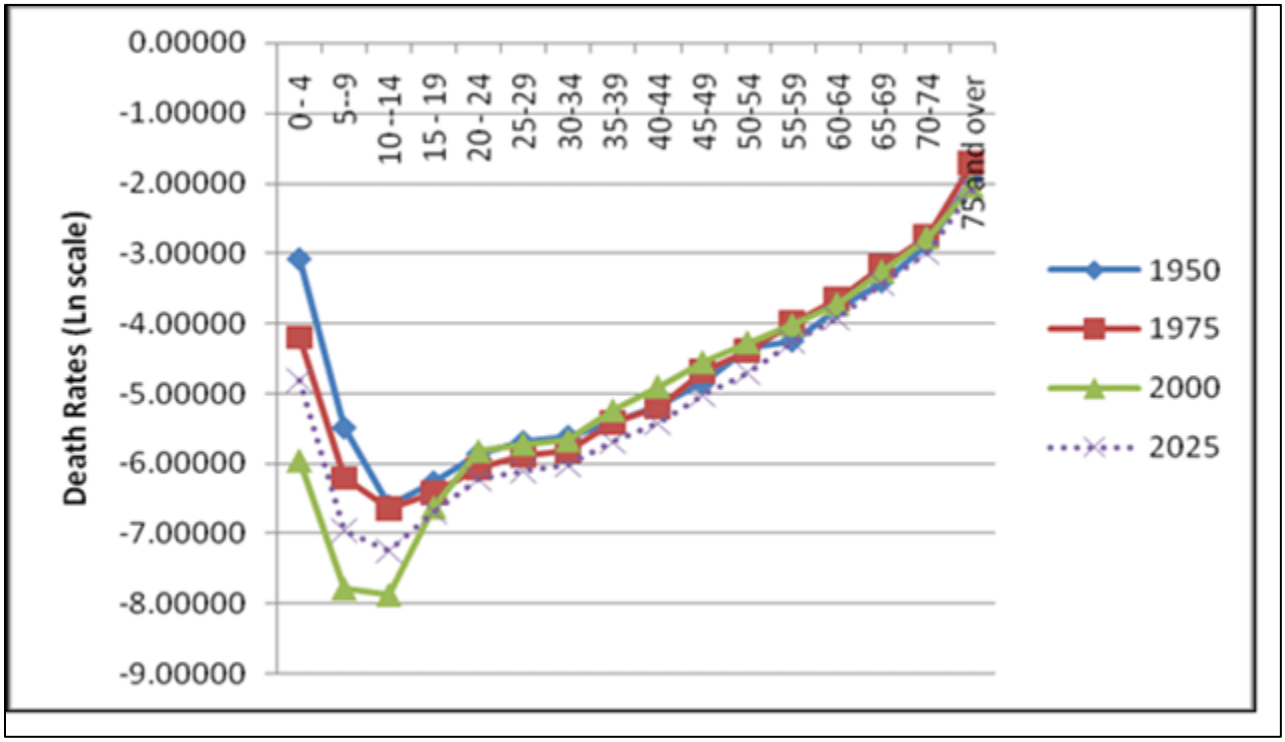

Figure 3.14: Survival to survival rate by 2025

More than 1.2 million people are killed in traffic accidents every year, and millions more are injured as a direct result of those incidents, as stated in the road safety technical package that was provided by the World Health Organization in 2017 [4] .As a direct consequence of this, the World Health Organization (WHO) presently ranks automobile accidents as the ninth most prevalent reason for people of any age to survive everywhere in the world [13], [14][15]. Statistics provided by the United Nations addressing the aging of populations around the world indicate that there were more than twice as many people aged 60 or older in 2017 as there were in 1980. In other words, the global population is becoming older. This rise can be linked to the fact that the average age of populations all across the world is always climbing higher. It is anticipated that the population of adults aged 60 and older across the globe would have more than doubled by the year 2050, reaching somewhere in the neighborhood of 2.1 billion individual [5].

It is going to be necessary for a number of countries to conduct tests and make changes to the resiliency of their transportation networks in order for those countries to be able to accommodate the influx of new inhabitants. This is something that will be expected of those countries. As the overall population continues to age, it is very likely that there will be a greater proportion of elderly people who require assistance when moving around. This is something that needs to be borne in mind at all times. Statistics were compiled by the National Highway Traffic Safety Administration for the year 2015, and it was discovered that the number of citizens who are over the age of 65 and hold a valid license to operate an automobile has significantly increased. This was one of the findings of the study. These numbers come from a poll that was taken in the year 2015, and they were derived from the responses received.

Elderly drivers have a lower likelihood of being involved in an automobile collision when compared to younger drivers; however, if they are involved in a collision, they have a higher risk of sustaining injuries. This is because elderly drivers tend to have weaker bones and muscles than younger drivers. The likelihood of a collision occurring for drivers under the age of 25 is significantly increased. A recent study came to the conclusion that drivers in the European Union have an increased risk of death in the event that they are involved in an accident with another vehicle on any of the routes that make up the union (EU). Incidents that are caused by drivers going at excessive speeds are a greater reason for concern for adults than collisions that are caused

by drivers who have insufficient experience behind the wheel. Drivers who travel at high speeds pose a larger risk to folks than other road users.

It is anticipated that in the not-too-distant future, global transportation networks will be confronted with a rising challenge when it comes to reducing the risk factors that are related with traffic accidents that are caused by drivers. A person whose job entails ensuring the safety of motorists and pedestrians is responsible, first and foremost, for identifying the factors that, when brought together, have the greatest potential for triggering an accident.

4. PROPOSED METHOD

4.1 PROPOSED METHOD

According to the findings of this study, a more sophisticated approach to assessing the chance of survival might prove to be quite useful in clinical situations. In addition, doctors who specialize in critical care can utilize this information to devise the best possible treatment strategy for their patients and communicate their findings to the patient's loved ones. Logistic regression (LR), random forest (RF), support vector machines (SVM) [18], XGBoost [20], and multilayer perceptron (MLP) [21-27] are just some of the statistical and AI models that have been put into use in clinical practice in recent years to assist in diagnostic suggestion [22], identify adverse events [23] and surgical complications [24], and predict in-hospital survival [25, 26]. In clinical practice, these models have been used to aid in diagnostic suggestion [22], identify Machine learning and fundamental statistical theories are used in today's high-performance computer systems in order to recognize and anticipate patterns. As can be seen in [26-29], the degree to which they are able to accurately forecast the future is contingent on the specifics of the situation. It is very new to use machine learning algorithms to the task of predicting whether or not a patient will survive a traumatic brain injury (TBI), and emergency rooms have not yet begun using this kind of technology to triage patients [30–34]. Administrators at the Chi Mei Center built a real-time prediction system for patients arriving to the emergency room with chest discomfort and significant adverse cardiac events by using the HIS's big-data driven methodology and machine learning capabilities [32–35]. In its current configuration, TTAS is able to identify the approach that will result in the fewest unnecessary re-triages and the most efficient use of available resources in the emergency department. The primary objective of this research was to establish the most effective strategy for making use of the TTAS in conjunction with several other distinguishing characteristics in order to forecast the possibility of an in-hospital mortality occurring among TBI patients. In this work, we developed a survival risk prediction model for traumatic brain injury patients with varied degrees of severity at first triage admission by applying machine learning algorithms to data from the HIS. This model was utilized to select patients for the study. This method was used to calculate the patient's likelihood of surviving after being admitted to the hospital. It is possible that an accurate method of prognostication might be created from the study of previous patients who were treated at the hospital for traumatic brain injuries. This strategy may

be utilized by professionals working in the medical area to assist them in determining the most appropriate course of therapy for their patients and to educate the loved ones of those patients.

4.1.1 Data Preparation

During the development process, there are a number of processes that need to be followed. These phases include selecting a learning method, defining the output variable, defining the input variables, and determining the model job. The selection of a preferred learning approach is one of the most important decisions that must be made while developing a prediction model. The proportional importance that we placed on validity and interpretability in our search for the greatest possible benefit had a role in the formation of our conclusion. For the better part of the 20th century, the models of logistic regression and proportional hazards were the workhorses of model development. As professionals become more experienced in working with the vast volumes of data that are accessible through electronic health records, they are increasingly turning to algorithmic approaches, which are more versatile than traditional statistical methods. These computational methods include deep learning and random forests, to name only two instances each.

Because they do not rely on a priori assumptions, these approaches are able to integrate more dimensions of information while simultaneously requiring fewer transformations and choices of variables than a regression-based approach (such as an assumption about the form of the function that relates the independent variables to the dependent variable).

During the development process, important stages include establishing the model task, collecting an adequate training sample for that task, defining the output variable, defining the input variables, and deciding on a learning technique. These are only a handful of the many responsibilities that need to be fulfilled. The process of establishing a trustworthy prediction model involves a number of phases, one of the most important of which is determining which instructional technique would produce the greatest outcomes. This choice was taken in large part on the basis of the relative priority that was given to the validity and interpretability of the results in order to maximize the value that could be derived from the findings of the investigation. Logistic regression and proportional hazards models have been utilized for the generation of the vast majority of models over the entirety of recorded history. There has been a movement toward employing algorithmic approaches, which are more pliable than more standard statistical methods. This change occurred

because algorithmic procedures are more efficient. This transformation has taken place about the same time as there has been an increase in access to the information contained in electronic medical records. A couple of examples of such computational methodologies are deep learning and the creation of random forests. These techniques can incorporate more dimensions of information with less need for transformations and choices of variables than a regression-based technique can be due to the lack of need for a priori assumptions, such as the shape of the function that links independent variables to a dependent variable. In other words, these techniques do not require a priori assumptions. It is possible to accomplish this using these methods since they need less adjustments.

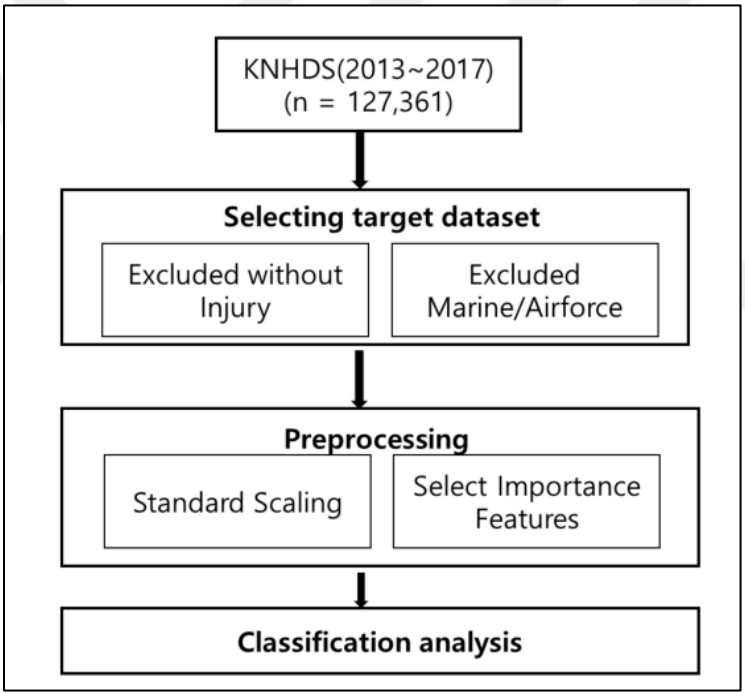

Figure 4.1: Outline of dataset preprocessing steps

4.1.2 Used Algorithms

The construction of five distinct categorization models in machine learning is covered in this article. These models were built with the use of five different popular supervised learning approaches. Our statistical toolkit is comprised of a variety of methodologies, a few examples of which are Logistic Regression, Decision Trees, Random Forests, and Naive Bayesian models. Comparisons are made about the capabilities of each model. The following concepts serve as the foundation for this and other algorithms in a similar vein: A certain type of algorithm known as

logistic regression may be utilized to make accurate projections on the results of an experiment (LR). A logistic function is used here in order to calculate the probabilities of each class occurring given the information that is currently available [[21]]. The LR calculation will tell you if this scenario has a zero percent chance of causing just minor injuries or a one hundred percent chance of causing fatalities (0). (1). At the core of the algorithm is a structure known as a decision tree.

The diagram that follows illustrates the fundamental structure of this approach, which includes the many nodes, branches, and leaves that comprise it. The node in this diagram represents a test variable, a single test result is represented by the branch, and a series of tests is represented by the leaf (i.e. no sever injury crash or Sever injury crash).

The path that a node takes from the root to the leaf determines whether it is considered a root or an internal node. The criteria for classification are established by the route that the node takes. When deciding whether or not to use the DT method, there are a number of benefits and downsides that should be considered, both of which will be investigated in further depth in the following sections. The implementation of this strategy demands a significant number of resources, despite the fact that it is successful when applied to smaller data sets [21].

We were able to generate five separate classification models by utilizing machine learning in our research. These models are produced using five distinct supervised learning algorithms that are often utilized in practice and are used in the generation process. With our assistance, you will have access to a wide variety of statistical approaches, including logistic regression, decision trees, random forests, and naive Bayesian models, amongst others. In addition, various methodologies, such as random forests and hierarchical linear models, are presented here for consideration. Logistic Regression and Random Forests are two more examples of the types of statistical procedures. In this post, we investigate and compare the capabilities of a number of different models. The following notions serve as the conceptual framework around which algorithms of this type are built: Logistic regression is a type of algorithm that may be employed in the process of making predictions about the findings of scientific study. This procedure can be carried out in relation to the results of various studies. The statistical method known as logistic regression (LR) may be utilized in the modeling of many types of connections between variables (LR). In this scenario, a logistic function is used to the data to reach an educated conclusion on the chance of each class occurring based on the information gathered [21]. As a consequence of this, it is now

able to produce forecasts that are more accurate. The proposed method architecture is presented in figure 4.3.

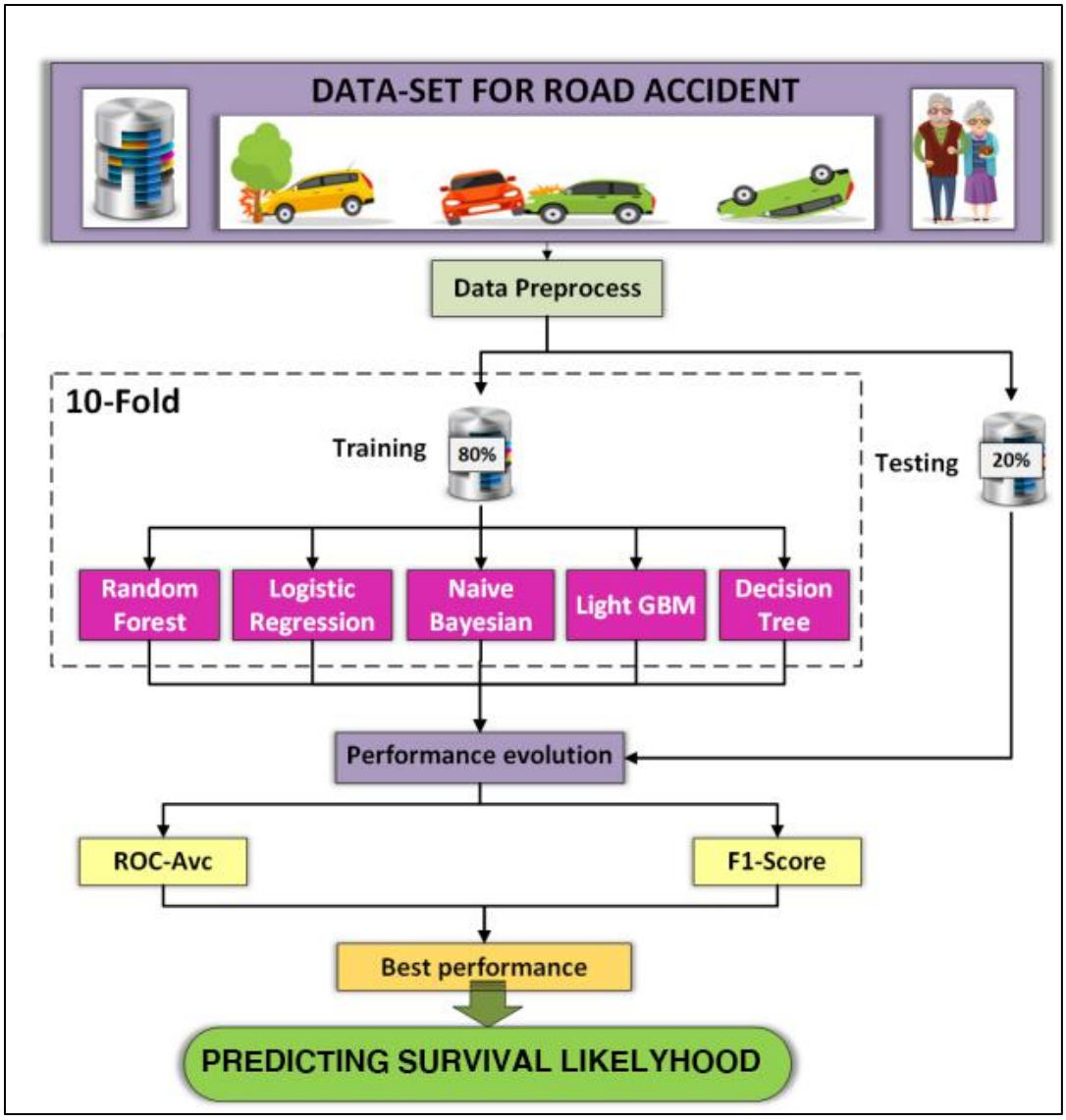

Figure 4.2: Proposed method architecture

i. The structure of the approach is mainly predicated on the use of a decision tree. Further down on this page, a figure is provided that illustrates the fundamental structure of this technique. The diagram shows the nodes, branches, and leaves that make up the method. Please scroll down the page a few notches in order to view the image. Thank you. As a result, a test variable is represented by a node, an individual test result is represented by a branch, and a group of tests is represented by a leaf (i.e. no sever injury crash or Sever injury crash). The path that a

node takes from the tree's root to its leaf defines whether or not that node is a root node or a leaf node. Along this path, the criteria for classification are determined. Before selecting whether or not to employ the DT technique, one must first consider its advantages and disadvantages in order to make an informed choice. In the next paragraphs, we'll proceed to investigate each of these subtopics in greater depth. When working with relatively small datasets, this strategy is efficient; nevertheless, when dealing with significantly larger datasets, it necessitates a significant amount of memory and processing time.

ii. Procedure using the Fast-GBM. An algorithm that is employed in the GBM is referred to as the "Gradient Boosting Machine," or GBM for short. In place of the more common depth-first strategy, this method implements a decision tree algorithm termed best first, which is a version of the usual technique. Using Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling, this method outperforms prior gradient-boosting frameworks in terms of training time and accuracy (EFB). The GOSS saves time and money on processing since it only evaluates a subset of the possible outcomes rather than all of them. The learning rate, maximum depth, number of leaves, and kind of boost [22] are just a few of the variables that affect the method's efficacy

iii. It's a piece of software known as the Random Forest algorithm (RF). The RF technique addresses the drawbacks of the DT methodology by constructing an arbitrary number of trees from the training dataset. This allows it to circumvent the issues described above. Every tree is allowed the opportunity to blossom to its utmost potential without any kind of pruning being done. The accuracy of a forecast may be determined by finding the mean of the forecasts made by a number of different trees [3]. The number of trees that were employed and the number of variables that were randomly sampled at each split are also displayed [2] since these are other parameters that influence RF.

iv. Algorithm known as the Naive Bayes (NB). This probabilistic approach is uncomplicated and simple to comprehend since it treats the characteristics of a class as if they were separate and unexplained things. A significant advantage of this method is that it is possible to estimate key parameters using only a limited sample of the available data [21].

4.2 RESULTS

The accuracy score of the models were built using the methods of logistic regression. The categorization models that were employed in this experiment are evaluated based on five different model assessment indicators, which are detailed in the study. The findings of the assessment revealed that performance was, on the whole, good, with scores ranging from -0.753 to +0.942. The accuracy score of 0.76 for models constructed using methodologies such as logistic regression and linear regression is lower than the industry average for these types of operations. The Rf algorithm came out on top with an accuracy score of 0.941, edging out the random forest approach, which finished in second place with a score of 0.844.

These models have a low F1 score, which indicates that they do not perform very well. The F1 score is a harmonic mean of accuracy and recall. In contrast, the F1 score for the random forest model was just 0.851, whereas the RF model's score was 0.942. When compared to other models found in the body of research, this one performs exceptionally well across the board in terms of the evaluation criteria. According to Figure 4.4, The models that were developed using the approaches of logistic regression and linear regression each had an accuracy score of 0.76, which is lower than the average accuracy score for models that are equal.

techniques utilized in the sector. The Rf algorithm came in first place with an accuracy score of 0.941, while the random forest technique came in second place with an accuracy score of 0.844. Both of these techniques were evaluated based on their ability to correctly predict outcomes. Because these models have a low F1 score, which is a harmonic mean of accuracy and memory, the performance of these models is expected to be below average. F1 is an acronym for the phrase "harmonic mean of accuracy and recall." The RF model achieved the highest F1 score of 0.942, which placed it in first place, while the random forest model achieved the lowest F1 score of 0.851, which placed it in second place. In terms of the evaluation criteria, this model performs far better than any of the other models that were looked into. The random forest model came in second place overall, and Figure 4 shows evaluation scores that are comparable to those given by the logistic regression model and the linear regression model respectively.

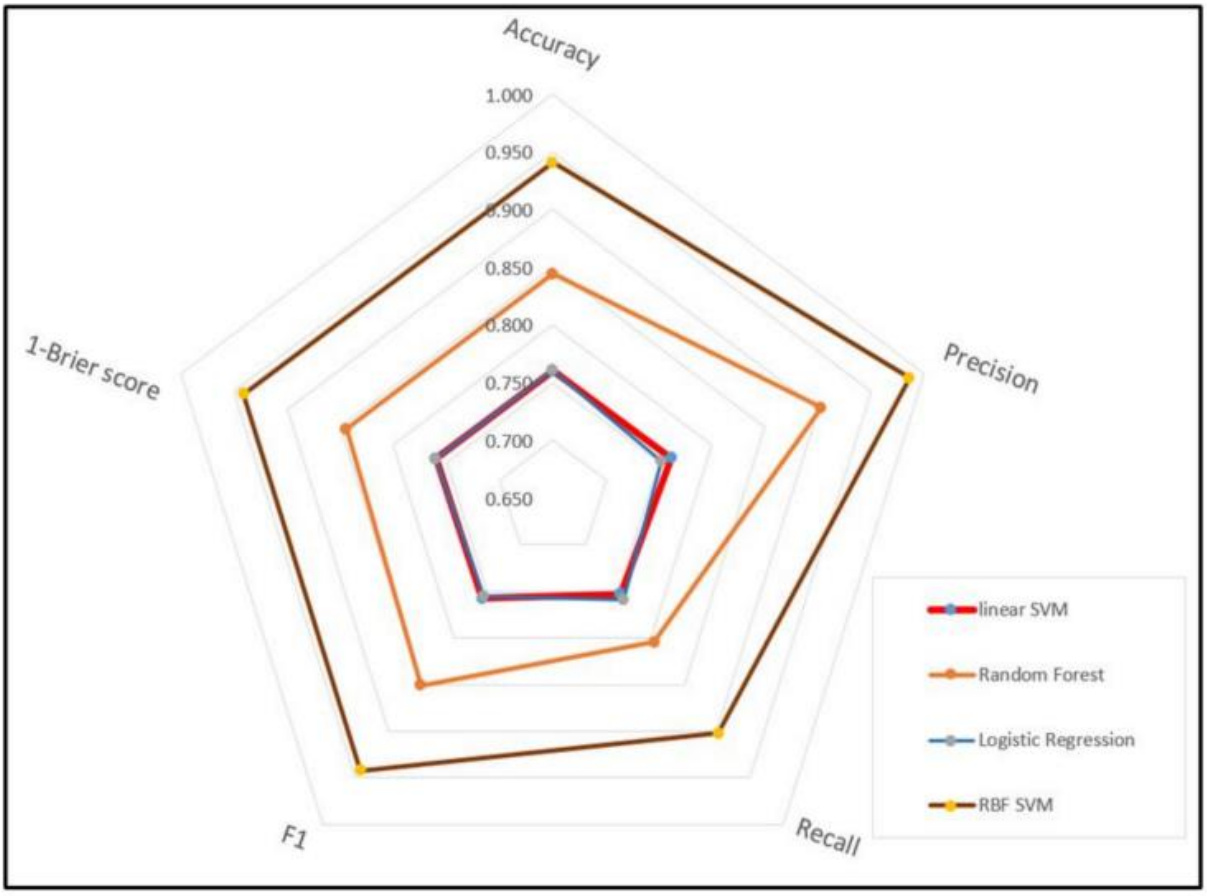

Figure 4.3: Evaluation of statistical analysis performance

In the validation phase, an improved model is fitted to the enhanced data and evaluated using the validation data and the under-sampling technique known as ten-fold cross-validation. In order to account for the stratification used in the original dataset, it was folded ten times, with the class distribution in each fold being quite close to that of the original. After transforming nine training samples into balanced data, the model was then fitted to the improved data. Afterward, the updated information was used to improve nine further training samples. After this is done, the process is repeated with the next set of subgroups. The final number is obtained by averaging the accuracy of each individual calculation as seen in figure 4.5.

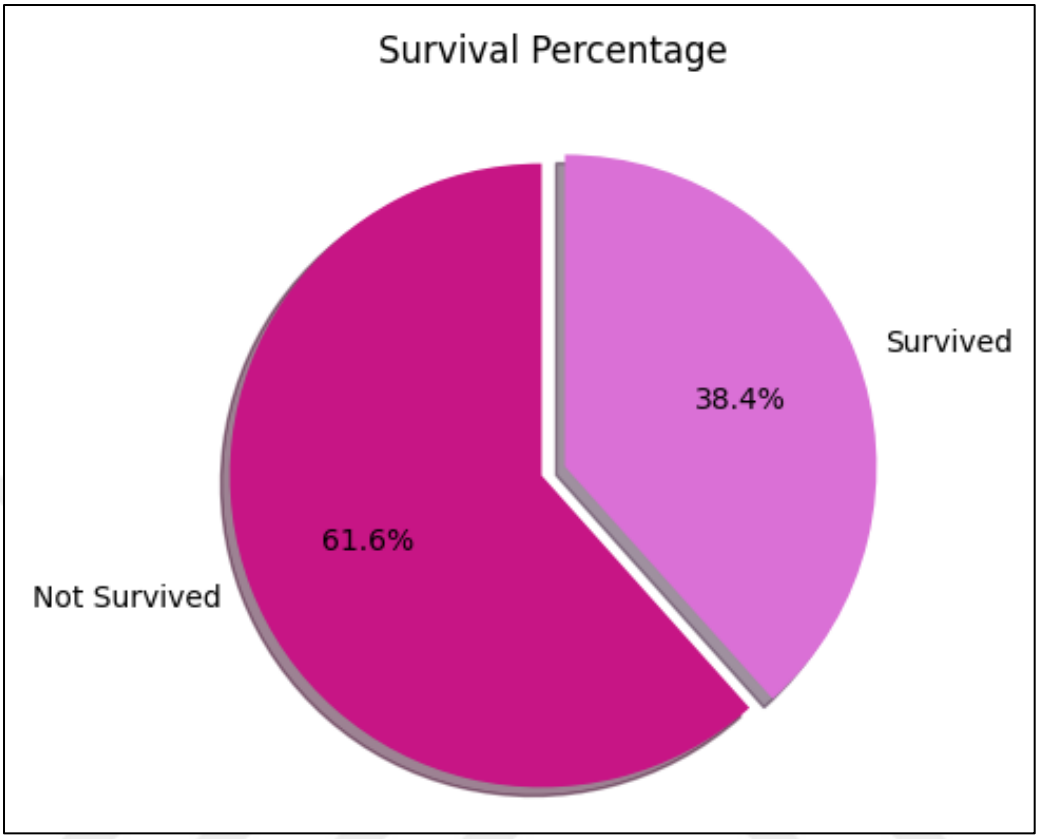

Figure 4.4: Rate of survivorship following verification of accuracy.

4.2.1 Result Discussion

In this chapter, we analyze accident data from the state of Michigan by making use of a number of different modeling approaches that are based on machine learning. These techniques include the NB, DT, LR, Light-GBM, and RF. A pre-processing procedure is utilized so that any missing data may be handled, and the data can be standardized. The information that was gathered was then used to produce a training set and a test set, which were both used in the subsequent analysis. We ensured that every individual in the community got access to the identical quantity of training data by utilizing the SMOTE approach. The models that are based on light-GBM data are proven to be the most accurate up to 87.97% of the time. It is possible to find out, with the assistance of the light-GBM model, the primary reason behind the mishaps that the elderly suffers from, which result in the most serious injuries.

In order to adopt a more proactive approach in the battle against traffic injuries among drivers in the state of Michigan, the Michigan Department of Transportation aims to employ this technology.

In addition, the data demonstrate that the average age of the population as well as the average daily traffic volume are the two aspects that are the most important to take into account. The increasing average age of both people and vehicles has had a significant impact on the changes that have taken place in this sector.

The results suggest that drivers should give considerable consideration to upgrading to newer models. Many more factors besides those listed above may be used in transportation sector research that employs deep learning to determine an older driver's risk score of traffic harm.

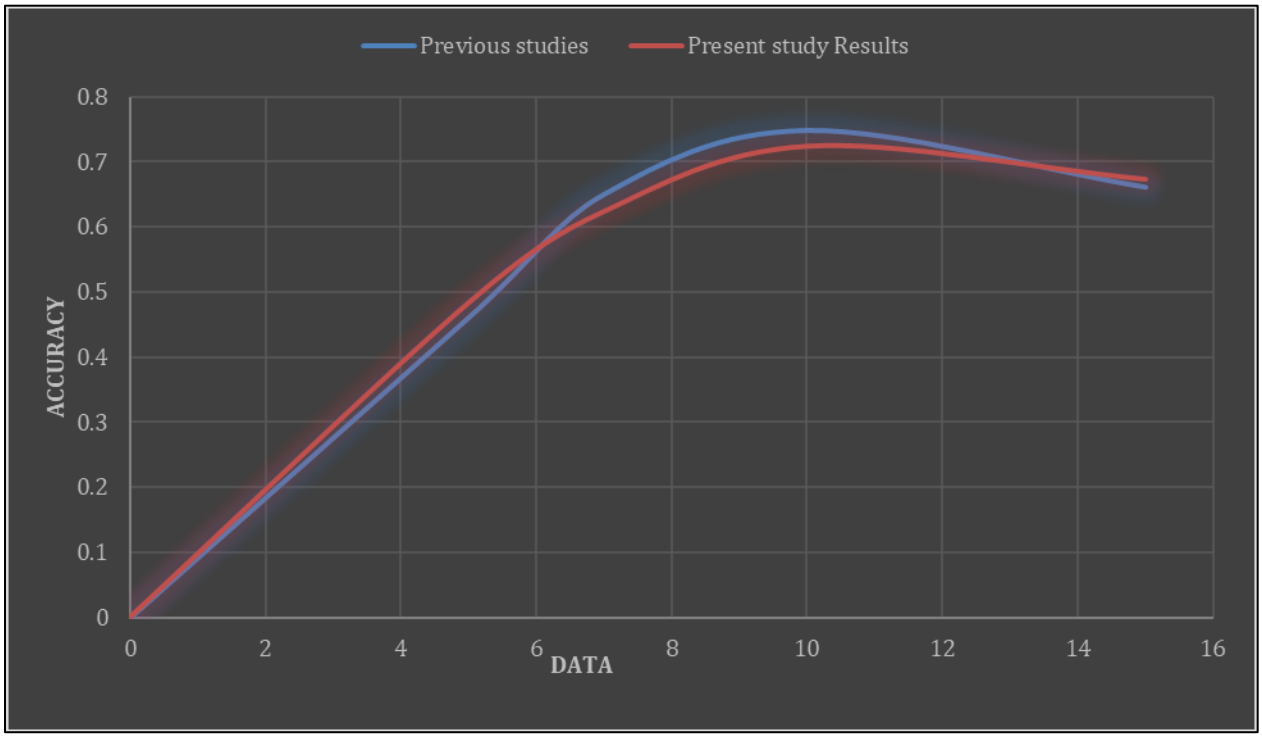

Figure 4.5: Results (accuracy, data)

In this chapter, classical survival prediction approaches have been compared with machine learning approaches. The classical metric of the MSE error on the force of death did not seem to be satisfactory, since it penalizes errors much more at high ages.

Due to sampling fluctuations, sometimes enormous coefficients of variation are obtained for ages above 95 years. This oriented us towards the choice of a new error metric (LMSE), as well as the choice to remove the points corresponding to ages above 95 years in the sample of data.

Table 4.1: Comparison of MSE results

Survival rate per 100 people	Previous studies	Present study Results	error%
0	5.7405E-05	0.0484375	0
5	10.0282793	10.53009558	5.00%
7	9.0378198	8.653446124	4.25%
10	7.35351947	7.299284202	0.74%
15	2.70469332	2.680605337	0.89%

It is interesting to make a link with the method of least squares which makes it possible to minimize the impact of experimental errors by “adding information” in the measurement process. If, as is usually the case, we have an estimate of the standard deviation of the noise that affects each measurement, we use it to “weigh” the contribution of the measurement: a measurement has all the more weight as its uncertainty is low. Applying this principle to our problem, it would be possible to create a new error metric (called a weighted MSE error) that would weight each point inversely to its associated variance. Both in the part dealing with the prediction of future survival and in the part discussing the modeling of a fictitious portfolio of annuities as a whole, the results obtained show that there is no There is no model that emerges in a clear way. Conventional survival models remain relevant and make it possible, for example, to easily obtain confidence intervals, machine learning approaches being particularly relevant in a multi-population framework (see the comparative report in section 4.4 for more details on the advantages and disadvantages of each approach).

In this chapter, we assess accident data from the state of Michigan using machine learning-based modeling approaches such as the NB, DT, LR, Light-GBM, and RF, amongst others. These modeling techniques were chosen because of their ability to predict the likelihood of certain types of accidents. The purpose of developing these models was to determine how likely it is that an accident will take place. These tactics were selected because of their capability to predict the consequences of accidents with a better degree of accuracy, which was a primary consideration in the decision-making process. In order to handle any missing data and normalize the data, a pre-

processing approach is utilized. This allows for the data to be standardized. This is done so that the data may be handled properly. After that, the information was used to construct two sets of data: a training set and a test set. The test set was used to evaluate the accuracy of the training set. The training set was utilized for educational purposes, whereas the test set was put to use for evaluation purposes. It was chosen to apply the SMOTE approach in order to make certain that the training data was spread uniformly throughout the population. This was accomplished by ensuring that the SMOTE methodology was used. This was done in order to ensure that our efforts were successful. Models that are data-based and employ light-GBM have the potential to attain the maximum level of accuracy, which can be as high as 87.97% of the time. When it comes to accidents involving elderly people that result in the most severe damage severity, the light-GBM model can be used to determine the underlying cause of the accident. This is the case even when it comes to incidents involving other age groups.

Because this methodology will be utilized by the Michigan Department of Transportation, it will make it possible for a more proactive approach to be taken toward reducing the number of injuries that are sustained by drivers in the state as a result of motor vehicle collisions. In addition, the data indicate that the age distribution of the population and the number of cars that travel through the area are the two aspects that ought to be considered the utmost importance. [Further citation is required] The rising average ages of both people and automobiles has been a significant element that has had a large impact on the development of the automotive industry. This trend can be seen in both developed and developing countries.

According to the results of the research, drivers over the age of 65 should give significant attention to the idea of purchasing a vehicle that was manufactured within the last ten years. A comparison study in the transportation sector using deep learning to compute the risk score of traffic injury to elder drivers has the potential to include a significant deal of additional pieces of information beyond those mentioned in the first portions of this article. This article has revealed a few of these further details.

5. CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

The accident data from the state of Michigan has been analyzed using a variety of machine learning-based modeling techniques, including the NB, DT, LR, Light-GBM, and RF models, amongst others. In order to handle data that is missing and to standardize the data, a preprocessing strategy is utilized. Following that, the information was put to use in the development of two sets of data: a training set and a test set. It was determined to adopt the SMOTE approach in order to guarantee that the training data was dispersed uniformly throughout the population. Up to 87.97% of the time, the most accurate models are those that are data-based and use light-GBM. When it comes to accidents involving elderly people that result in the most severe damage severity, the light-GBM model can be used to determine the underlying cause of the accident. This methodology, which will be utilized by the Michigan Department of Transportation, will make it possible for a more proactive approach to be taken toward minimizing the number of traffic injuries sustained by senior drivers in the state.

In addition, the data indicate that the age distribution of the population and the number of vehicles moving through the area are the two most significant aspects to take into account. The evolution of the automotive industry has been significantly influenced by the advancing ages of both people and cars. The research indicates that motorists over the age of 65 should think about purchasing more contemporary automobiles. A comparative research study in the transportation business that uses deep learning to build the risk score of traffic injury to elderly drivers has the potential to include a great deal of other components, in addition to those that have been presented thus far.

5.2 FUTURE WORK

Home security systems such as Netatmo, Netgear, Honeywell, and Ooma all make use of face recognition technology. Because of this, the systems are able to identify individuals who are coming close to the property and to discover potential invaders. This helps to prevent unauthorized entry into the systems in question. Its potency and significance are null and void when there is no one there to see it and interpret it. Honeywell, as part of an effort to partner with Amazon Alexa, has built and introduced a groundbreaking new home security system that is now available to

customers. Customers can purchase this system right now. This system, which was conceived of and created by Honeywell, is perfect for people who want to build a smarter house in a variety of ways, and it is excellent for those who fit into these categories. Honeywell devised and developed this system. This system was constructed and developed by Honeywell.

At this time, the Google Nest Cam IQ is able to identify people from a distance of up to fifty meters away, and it can also maintain a vigilant eye on your property at any time of the day or night. Both of these capabilities are already available. At this time, the United States of America is the only country to offer this capability. In the event that it finds a living being in the vicinity, the security system can be programmed to open either the main door or the gate, depending on which one was designated to do so in the initial configuration. Utilizing code will allow for the successful completion of this task.

Even though it has the potential to have a significant impact on our lives in the future, face recognition already has an effect on how technology addresses our safety and other aspects of our lives [28–36]. This is the case despite the fact that face recognition already has an effect on these aspects of our lives. The current state of affairs has not changed despite the fact that advances in technology have the potential to bring about huge changes in the manner in which we live our lives. Customers now have access to a vast amount of information that has been arranged in a variety of categories thanks to the proliferation of the World Wide Web. The data suggest that Deep Learning clusters frequently perform better than their rivals, which was a topic that was brought up in conversation with the client and then refined together.

It is feasible to improve the performance of a cluster by obtaining information from other clusters that are currently operating at a better level of performance and implementing that information into the cluster in question. It was asserted that using the application will cut down not just on the amount of time spent learning but also on the weight that is computed for the network. This was one of the claims that was made. In the Management cluster, an increase in the learning coefficient is related with a fall in the overall quality of the cluster, whereas a decrease in the learning coefficient is associated with an improvement in the overall quality of the cluster when it has a lower value. This association is still valid in the event that the learning coefficient is interpreted as being negative. It is necessary to do a precise calibration on the Management Cluster in order for it to be able to produce the outcomes that are desired.

REFERENCES

[1] G. Xiao, J. Guo, L. da Xu, and Z. Gong, "User interoperability with heterogeneous IoT devices through transformation," *IEEE Trans Industr Inform*, vol. 10, no. 2, pp. 1486–1496, 2014, doi: 10.1109/TII.2014.2306772.

[2] L. Chettri and R. Bera, "A Comprehensive Survey on Internet of Things (IoT) Toward 5G Wireless Systems," *IEEE Internet Things J*, vol. 7, no. 1, pp. 16–32, Jan. 2020, doi: 10.1109/JIOT.2019.2948888.

[3] F. Viani, F. Robol, A. Polo, P. Rocca, G. Oliveri, and A. Massa, "Wireless architectures for heterogeneous sensing in smart home applications: Concepts and real implementation," *Proceedings of the IEEE*, vol. 101, no. 11, pp. 2381–2396, 2013, doi: 10.1109/JPROC.2013.2266858.

[4] A. Geetha and M. Dol, "A Learning Transition from Machine Learning to Deep Learning: A Survey", doi: 10.1109/ICETCI51973.2021.9574066.

[5] F. Chollet, "Deep Learning with Python, Second Edition," *Deep Learning with Python*, 2021, Accessed: Nov. 28, 2022. [Online]. Available: https://www.manning.com/books/deep-learning-with-python-second-edition

[6] L. Liu *et al.*, "Deep Learning for Generic Object Detection: A Survey," *Int J Comput Vis*, vol. 128, no. 2, pp. 261–318, Feb. 2020, doi: 10.1007/S11263-019-01247-4/FIGURES/21.

[7] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature 2015 521:7553*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," *undefined*, 2011.

[9] V. Kepuska and G. Bohouta, "Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home)," *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018*, vol. 2018-January, pp. 99–103, Feb. 2018, doi: 10.1109/CCWC.2018.8301638.

[10] E. Strickland, “IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care,” *IEEE Spectr*, vol. 56, no. 4, pp. 24–31, Apr. 2019, doi: 10.1109/MSPEC.2019.8678513.

[11] S. A. Bini, “Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?,” *J Arthroplasty*, vol. 33, no. 8, pp. 2358–2361, Aug. 2018, doi: 10.1016/J.ARTH.2018.02.067.

[12] B. van Ginneken, “Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning,” *Radiol Phys Technol*, vol. 10, no. 1, pp. 23–32, Mar. 2017, doi: 10.1007/S12194-017-0394-5/FIGURES/2.

[13] D. Ravi *et al.*, “Deep Learning for Health Informatics,” *IEEE J Biomed Health Inform*, vol. 21, no. 1, pp. 4–21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665.

[14] R. Hecht-Nielsen, “Theory of the backpropagation neural network,” pp. 593–605, 1989, doi: 10.1109/IJCNN.1989.118638.

[15] J. R. Zhang, J. Zhang, T. M. Lok, and M. R. Lyu, “A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training,” *Appl Math Comput*, vol. 185, no. 2, pp. 1026–1037, Feb. 2007, doi: 10.1016/J.AMC.2006.07.025.

[16] M. J. Syu and B. C. Chen, “Back-propagation Neural Network Adaptive Control of a Continuous Wastewater Treatment Process,” *undefined*, vol. 37, no. 9, pp. 3625–3630, Sep. 1998, doi: 10.1021/IE9801655.

[17] A. Jalal and S. Kamal, “Real-time life logging via a depth silhouette-based human activity recognition system for smart home services,” *11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2014*, pp. 74–80, Oct. 2014, doi: 10.1109/AVSS.2014.6918647.

[18] S. Chen, S. Ding, H. Fu, Y. Xian, X. Liu, and C. Zhang, “Deep Learning Applied to Smart Home Face Recognition Access Control System,” pp. 13–15, Mar. 2018, doi: 10.2991/ICAITA-18.2018.4.

[19] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning.” PMLR, pp. 1139–1147, May 26, 2013. Accessed: Nov. 28, 2022. [Online]. Available: https://proceedings.mlr.press/v28/sutskever13.html

[20] S. Pawar, V. Kithani, S. Ahuja, and S. Sahu, “Smart Home Security Using IoT and Face Recognition,” *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, Jul. 2018, doi: 10.1109/ICCUBEA.2018.8697695.

[21] D. A. R. Wati and D. Abadianto, “Design of face detection and recognition system for smart home security application,” *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, vol. 2018-January, pp. 342–347, Feb. 2018, doi: 10.1109/ICITISEE.2017.8285524.

[22] M. Mathew and R. S. DIvya, “Super secure door lock system for critical zones,” *2017 International Conference on Networks and Advances in Computational Technologies, NetACT 2017*, pp. 242–245, Oct. 2017, doi: 10.1109/NETACT.2017.8076773.

[23] S. Anwar and D. Kishore, “IOT based Smart Home Security System with Alert and Door Access Control using Smart Phone,” *undefined*, vol. V5, no. 12, Dec. 2016, doi: 10.17577/IJERTV5IS120325.

[24] F. Zuo and P. H. N. de With, “Real-time embedded face recognition for smart home,” *IEEE Transactions on Consumer Electronics*, vol. 51, no. 1, pp. 183–190, Feb. 2005, doi: 10.1109/TCE.2005.1405718.

[25] S. Yi, Z. Hao, Z. Qin, and Q. Li, “Fog computing: Platform and applications,” *Proceedings - 3rd Workshop on Hot Topics in Web Systems and Technologies, HotWeb 2015*, pp. 73–78, Jan. 2016, doi: 10.1109/HOTWEB.2015.22.

[26] O. Patsadu, C. Nukoolkit, and B. Watanapa, “Human gesture recognition using Kinect camera,” *JCSSE 2012 - 9th International Joint Conference on Computer Science and Software Engineering*, pp. 28–32, 2012, doi: 10.1109/JCSSE.2012.6261920.

[27] V. Sati, D. Garg, T. Choudhury, and A. Aggarwal, “Facial recognition-application and future: A review,” *Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends, SMART 2018*, pp. 231–235, Nov. 2018, doi: 10.1109/SYSMART.2018.8746942.

[28] K. Kim, J. Lee, and W. Jung, “Method of building a security vulnerability information collection and management system for analyzing the security vulnerabilities of IoT devices,” *Lecture Notes in Electrical Engineering*, vol. 448, pp. 205–210, 2017, doi: 10.1007/978-981-10-5041-1_35.

[29] J. D. McLean, C. Herley, and P. C. van Oorschot, “Letter to the Editor,” *IEEE Secur Priv*, vol. 16, no. 03, pp. 6–10, May 2018, doi: 10.1109/MSP.2018.2701158.

[30] “Build Security In Maturity Model (BSIMM) – Practices from Seventy Eight Organizations.” https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=450642 (accessed Nov. 28, 2022).

[31] M. Park, H. Oh, and K. Lee, “Security Risk Measurement for Information Leakage in IoT-Based Smart Homes from a Situational Awareness Perspective,” *Sensors (Basel)*, vol. 19, no. 9, May 2019, doi: 10.3390/S19092148.

[32] N. Romano, “Securing Our Future Homes: Smart Home Security Issues and Solutions,” *Honors Theses*, Apr. 2019, Accessed: Nov. 28, 2022. [Online]. Available: https://digitalcommons.liberty.edu/honors/864

[33] J. Bugeja, A. Jacobsson, and P. Davidsson, “On privacy and security challenges in smart connected homes,” *Proceedings - 2016 European Intelligence and Security Informatics Conference, EISIC 2016*, pp. 172–175, Mar. 2017, doi: 10.1109/EISIC.2016.044.

[34] M. Abdur, S. Habib, M. Ali, and S. Ullah, “Security Issues in the Internet of Things (IoT): A Comprehensive Study,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017, doi: 10.14569/IJACSA.2017.080650.

[35] K. Ghirardello, C. Maple, D. Ng, and P. Kearney, “Cyber security of smart homes: Development of a reference architecture for attack surface analysis,” *IET Conference Publications*, vol. 2018, no. CP740, 2018, doi: 10.1049/CP.2018.0045.

[36] A. Severyn and A. Moschittiy, “Learning to rank short text pairs with convolutional deep neural networks,” *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 373–382, Aug. 2015, doi: 10.1145/2766462.2767738.

[37] L. Jiang, D. Wang, Z. Cai, and X. Yan, “Survey of improving Naive Bayes for classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4632 LNAI, pp. 134–145, 2007, doi: 10.1007/978-3-540-73871-8_14/COVER.

[38] L. Jiang, Z. Cai, D. Wang, and S. Jiang, “Survey of improving K-nearest-neighbor for classification,” *Proceedings - Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007*, vol. 1, pp. 679–683, 2007, doi: 10.1109/FSKD.2007.552.

[39] P. A. Flach and N. Lachiche, “Naive Bayesian Classification of Structured Data,” *Machine Learning 2004 57:3*, vol. 57, no. 3, pp. 233–269, Dec. 2004, doi: 10.1023/B:MACH.0000039778.69032.AB.

[40] L. E. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: 10.4249/SCHOLARPEDIA.1883.

[41] L. Rokach and O. Maimon, “Decision Trees,” *Data Mining and Knowledge Discovery Handbook*, pp. 165–192, May 2006, doi: 10.1007/0-387-25465-X_9.

[42] Q. Zhang, Y. Yang, H. Ma, Y. N. Wu, and † Shanghai, “Interpreting CNNs via Decision Trees.” pp. 6261–6270, 2019.

[43] Z. ÇETİNKAYA and F. HORASAN, “Decision Trees in Large Data Sets,” *International Journal of Engineering Research and Development*, vol. 13, no. 1, pp. 140–151, Jan. 2021, doi: 10.29137/UMAGD.763490.

[44] A. Salimi, R. S. Faradonbeh, M. Monjezi, and C. Moormann, “TBM performance estimation using a classification and regression tree (CART) technique,” *Bulletin of Engineering Geology and the Environment*, vol. 77, no. 1, pp. 429–440, Feb. 2018, doi: 10.1007/S10064-016-0969-0/FIGURES/7.

[45] “Decision Trees: ID3 Algorithm Explained | Towards Data Science.” https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1 (accessed Nov. 30, 2022).

[46] O. U. Biometrics, A. Shakiru, and R. R. Hocking, “A Biometrics invited paper. The analysis and selection of variables in linear regression On Model Select ion Crit eria of Mult ivariat e Ridge Regression and Ordinary Least Square Regression A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression,” *Biometrics*, vol. 32, no. 1, pp. 1–49, 1976.

[47] K. Kumari and S. Yadav, “Linear regression analysis study,” *Journal of the Practice of Cardiovascular Sciences*, vol. 4, no. 1, p. 33, 2018, doi: 10.4103/JPCS.JPCS_8_18.

[48] A. Zakutayev *et al.*, “An open experimental database for exploring inorganic materials,” *Scientific Data 2018 5:1*, vol. 5, no. 1, pp. 1–12, Apr. 2018, doi: 10.1038/sdata.2018.53.

[49] M. Aykol, S. S. Dwaraknath, W. Sun, and K. A. Persson, “Thermodynamic limit for synthesis of metastable inorganic materials,” *Sci Adv*, vol. 4, no. 4, Apr. 2018, doi: 10.1126/SCIADV.AAQ0148/SUPPL_FILE/AAQ0148_SM.PDF.

[50] J. Shen *et al.*, “Reflections on one million compounds in the open quantum materials database (OQMD),” *Journal of Physics: Materials*, vol. 5, no. 3, p. 031001, Jul. 2022, doi: 10.1088/2515-7639/AC7BA9.

[51] S. Kirklin *et al.*, “The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies,” *npj Computational Materials 2015 1:1*, vol. 1, no. 1, pp. 1–15, Dec. 2015, doi: 10.1038/npjcompumats.2015.10.