POLITECNICO DI MILANO

Bioingegneria Elettronica ed Informatica

Automatic COmputation of cardioVascular arrhythmIc risk from ECG Data of COVID 19 patients- COVIDSQUARED

TESINA LAURA MAGISTRALE IN INGEGNERIA BIOMEDICA

Author: Naeime Saeidi

Student Id:10641463

Advisor: Prof. Luca Mainardi

Academic year: 2021-2022

Contents

List of figures

List of tables

Abstract

Heart disease is one of the leading causes of death in recent years, and this issue has increased the importance of early and rapid diagnosis of cardiac arrhythmias. Cardiac arrhythmias are a group of conditions in which the electrical activity of the heart is irregular and faster or slower than normal.

An electrocardiogram (ECG) is widely used as a simple and non-invasive way to diagnose heart disease, especially dangerous and common arrhythmias. In this project, we intend to design a system that can automatic computation of cardiovascular arrhythmic risk from ECG of covid 19 patients.

Using machine learning techniques to automatically detect and parse these signals can be very helpful. In general, ECG analysis is performed in two stages: the first stage is the extraction of ECG properties and the second stage is their classification based on the extracted characteristics to different conditions. Manual analysis is very time consuming and tedious, so the development of automated methods for ECG analysis is crucial. In recent years, many algorithms have been proposed to detect cardiac arrhythmias, which mainly include four main methods. These methods include noise cancellation, waveform detection, feature extraction, and arrhythmia classification.

Chapter 1:
Introduction and basic concept

Introduction

Heart disease is one of the leading causes of death in recent years, and this issue has increased the importance of early and rapid diagnosis of cardiac arrhythmias. Cardiac arrhythmias are a group of conditions in which the electrical activity of the heart is irregular and faster or slower than normal. Cardiac arrhythmias are caused by heart diseases such as myocardial infarction, cardiomyopathy, and myocardial infarction and are associated with a risk of death for patients[1]. Cardiac arrhythmias affect more than 4 million people in the United States and cost the health care sector up to $ 67.4 billion annually, which is an economically significant burden. Cardiac arrhythmias can also have many complications, such as an increased risk of stroke and death[2]. The development of methods and tools for diagnosing heart disease is one of the most important debates in the field of medical engineering. And the use of ECG signals in this field is very important. Cardiac arrhythmias are associated with abnormal electrical activity in the heart that can be reflected by ECGs[1]. ECG analysis is one of the most common methods in diagnosing cardiac arrhythmias due to its simplicity and low cost. And huge amounts of ECG data are collected daily in hospitals and homes, and this high volume may prevent them from being carefully examined and analyzed by human operators. Using machine learning techniques to automatically detect and parse these signals can be very helpful. In general, ECG analysis is performed in two stages: the first stage is the extraction of ECG properties and the second stage is their classification based on the extracted characteristics to different conditions. Manual analysis is very time consuming and tedious, so the development of automated methods for ECG analysis is crucial. In recent years, many algorithms have been proposed to detect cardiac arrhythmias, which mainly include four main methods. These methods include noise cancellation, waveform detection, feature extraction, and arrhythmic classification[1].

Coronavirus 2019 (COVID-19) is a disease caused by infection with Acute Respiratory Syndrome-Coronavirus-2 (SARS-CoV-2), which is known as a public health emergency and has caused the current epidemic in the world. Facing the challenges posed by the virus requires immediate and precise action. The virus uses the converter enzyme 2 (ACE2) as a functional receptor to enter the cell, and this protein plays a role in heart function and the pathophysiology of diabetes mellitus and hypertension. The SARS-CoV-2 virus can also use this entry route as a route to attack and directly damage myocardial cells[3].

Although the clinical signs and symptoms of Covid 19 are predominantly respiratory; Observations of major cardiac complications have also been reported. The cause of cardiac complications appears to be multifactorial, including direct viral myocardial damage, hypoxia, hypotension, increased inflammatory status, decreased ACE2 receptors, drug poisoning, adrenergic status of internal catecholamines, and others. Preliminary studies show that coronavirus disease is associated with an increased incidence of cardiac arrhythmias. Covid 19 disease may damage the myocardium and increase the risk of arrhythmias. Preliminary reports from China indicate that cardiac arrhythmias occur in 17% of hospitalized patients. A heart rate of 44% is higher in patients admitted to the intensive care unit (ICU).

Recent studies also show that myocardial injury is particularly common in patients with severe cuvitis.

In this study, data from patients with Quid 19 including electrocardiographic outputs and personal information such as age and sex were collected by studying and analyzing these data to provide an automated method for calculating the risk of cardiac arrhythmias in these individuals, using methods based on We focus on data mining and machine learning algorithms. For this reason, in this chapter, the definition and importance of the subject will be discussed first and the necessity of research will be stated. Then, a brief definition of the basic concepts used in the research will be provided.

Subject definition

Due to the outbreak of coronary heart disease and its epidemic around the world, it has created problems and problems for human beings that have caused financial and human damage to them. For this reason, due to the advancement of technology and with the emergence of smart and data-driven organizations, relevant services have become very important to meet the needs of this issue. To this end, a lot of data is generated in different types for these types of issues. These data are sometimes so complex that they cannot be analyzed by traditional methods. For this reason, newer methods such as data mining and machine learning must be used to obtain more accurate answers and reduce injuries. Therefore, in this research, data mining and machine learning will be used to solve the problem.

Importance and necessity of the subject

Since 2019, when the epidemic of coronary heart disease, as a respiratory disease, brought with it many problems. For this reason, it attracted the attention of researchers around the world. It has also had a tremendous impact on the social and economic situation of countries. The importance of this issue is to the extent that countries are facing many crises for reasons such as high mortality of citizens. For this reason, due to the epidemic and the referral of many people to medical centers, a lot of data has been generated daily and has caused problems. Problems such as predicting patient mortality, predicting cardiovascular and respiratory health, deciding whether to prescribe the appropriate medication for patients, and choosing the appropriate vaccine for patients. To solve these problems, various methods can be used in which new methods and techniques are superior. In other words, methods such as data mining, data analysis and machine learning can be very helpful.

Research purposes

The main purpose of this study is to conduct a systematic and comprehensive literature review to find different areas for solving problems related to coronary heart disease and its injuries, as well as the application of various data mining and machine learning methods based on research conducted in this field. Also, choosing the appropriate method to be used in the data set and solving the problem is another goal of this research. In fact, in summary, the objectives of this research can be stated in the following cases:

- Identify areas related to solving coronary heart disease problems
- Application of data mining techniques in various fields of health and field related to Covid disease 19
- Presenting the appropriate method and solving the desired problem and evaluating the results obtained from the method

Basic definitions of research

Heart

The heart is a muscle organ in humans and other animals that circulates blood through the blood vessels in the circulatory system. Blood provides oxygen and nutrients to the body and also helps eliminate waste products from metabolism.

The human heart beats at an average of about 70 beats per minute at rest. The human heart is located between the two lungs in the body and is positioned so that its head is tilted to the left and down. Each heartbeat takes about eight tenths of a second, which includes 0.1 seconds of atrial contraction, 0.3 seconds of ventricular contraction, and 0.4 seconds of rest of the heart. Heart tissue, like other tissues in the body, needs nourishment. The heart is nourished by the coronary arteries. The heart is located in the middle space of the middle of the chest.[4]

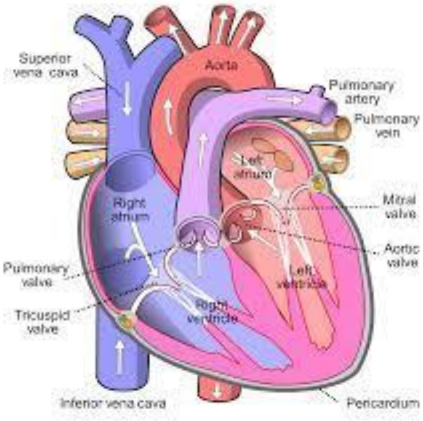


Figure 1: Heart

The word heart means transformation. The heart is scientifically an organ located in the center of the chest (inclined to the left) and weighs 300 grams and pumps blood. The name of this organ is because it changes the blood and turns dirty blood into The blood is cleansed. The heart is registered. In transforming the blood, the heart acts like a pump.

In humans, other mammals and birds of the heart are divided into four cavities: the left and right atria, which are above, and the left and right ventricles, which are located at the bottom. [5] The right atrium and ventricle are usually known as the right heart and their left counterparts as the left heart. But in other cases, the heart of fish has two chambers, one ventricle and one atrium, while the heart of reptiles has three chambers. In a healthy heart, due to the presence of heart valves, blood enters the heart from one side, which prevents the reverse flow. The heart is housed in a protective sac called the pericardium, which also contains fluid. The wall of the heart is made up of four layers that stick together, from the outside, including the pericardium (epicardium), the heart muscle (myocardium), and the endocardium (endocardium). This conical organ, in the form of a muscular sac, is located approximately in the middle of the thoracic space (middle mediastinum) slightly forward and to the left, between the left and right lungs, and is inclined to the left lung, which causes the shape and placement of the lungs to be different. Because the heart is a very sensitive and vital organ, it is protected by the thorax. The size of the heart in an adult is about 6 by 9 by 12 centimeters and its mass in men is about 300 and in women, about 250 grams, ie about 0.4 percent of the total body mass.

The two lower chambers of the heart are called the ventricles. The heart has a left ventricle and a right ventricle. In the middle of the heart, there is a thick muscle wall between the two ventricles called the septum. The septum works by separating the right side of the heart from the left side of the heart.

The ventricles are the two lower chambers of the heart that are separated by thick, strong muscle walls. The size of the ventricles is larger than the atria; And their job is to pump blood out of the heart.

The right ventricle receives blood through the right atrium through the right atrium; And then sends it through the pulmonary valve to the pulmonary artery and to the lungs. The left ventricle receives oxygenated blood from the left atrium through the mitral valve (bivalve) and sends it through the aortic valve to the aorta and thus to all tissues of the body.

The walls of the ventricles are thicker than the walls of the atria. The walls of the ventricles are thicker than the walls of the atria, because the blood pressure that flows into or out of the atria is much lower than blood pressure from the ventricles into the arteries (aorta and pulmonary artery). Becomes; Therefore, the diameter and strength of the ventricular walls make their resistance to this pressure possible.

Cardiac arrhythmia

Arrhythmia is an abnormal heart rhythm. This may be just a temporary pause and be so short that it does not affect the overall heart rate, or it may cause the heart to beat too fast or too slow. Some arrhythmias do not cause any symptoms. Other arrhythmias may cause symptoms such as lightheadedness or dizziness. There are two main types of arrhythmia: Bradycardia occurs when the heart rate is very slow (less than 60 beats per minute). Tachycardia also occurs when the heart rate is very fast (more than 100 beats per minute).

If the arrhythmia is short, it usually has no symptoms. It may just be in the form of missing a heartbeat that you rarely notice. There may be a sensation of tremor (vibration) in the chest or neck.

When the arrhythmias are severe or last long enough to affect heart function, the heart may not be able to pump enough blood to the body. This can lead to feelings of tiredness or lightheadedness or lethargy. It can also lead to death.

Tachycardia can lead to decreased ability to pump the heart, resulting in shortness of breath, chest pain, lightheadedness, or decreased level of consciousness. This condition can lead to a heart attack or death if severe.

Electrocardiography

Electrocardiography is the process of electrocardiogram, which is a recorded diagram of changes in electrical potential caused by stimulation of the heart muscle. It is usually identified by the abbreviation ECG or EKG (the latter case stands for the German word Elektrokardiogramm).

The electrocardiograph continuously records this diagram on a dedicated striped paper tape. The information recorded on the electrocardiogram shows the electrical waves of the heart. These waves indicate the different stages of cardiac stimulation. The curve that is drawn is called an "electrocardiogram." Doctors can use this curve to see how the heart works. Each curve contains three waves. The p-wave shows shortly before the electrical activity of the atria, the QRS complex shortly before the electrical activity of the ventricles, and the T-wave represents the rest of the ventricles[6].

At this stage, the ventricles and atria of the heart are at rest. Dark blood flows into the right atrium through the great superior and inferior vena cava (veins of the upper and lower cavities). Due to its weight, this blood enters the ventricles through the atrioventricular valves, which open at the end of the previous T-wave of the cardiac cycle, and fills them to some extent. But in order for atrial blood to enter the ventricle completely, the atria must contract. It should be noted that any muscle in the heart that wants to contract or relax must first propagate its contraction or relaxation wave to all parts of that muscle. Therefore, in order for the atria to contract, the contraction message must first be propagated throughout them. This is done by atrial node tissue. Between the two atria, this is the only right atrium that has nodular tissue. On the other hand, the heart's contraction, which is the node, is located in the posterior wall of the right atrium and under the large opening of the superior vena cava. For contraction, then, the precursor node is stimulated spontaneously in a rhythm, and this message conducts the contraction through the three strings of the right atrial node to the atrioventricular node, which is located between the atrial wall and the ventricle and is slightly inclined to the right atrium.

As the message travels from the scout to the atrioventricular septum, the myocardial ions of the heart that carry the message contract, which propagate from ion to ion in the right atrium and eventually through the atria to the left atrium. And covers the entire atria.

Of course, this message cannot be transmitted through the atria to the ventricles, because in the wall between the ventricles and atria is the connective tissue of the insulated fibers, which allows the message to be transmitted from the atria to the ventricles only through the nodal tissue that passes through this insulation. . If this tissue were not insulated, the atria and ventricles would contract at the same time and heart function would be very low; Because in this case, after pumping a small amount of blood to the ventricles, they also pumped the same small amount to the body and lungs and a little blood reached them.

The P-wave is recorded on the electrocardiogram after the message has completely covered the entire atrium. Immediately after that, the general rest period of the heart, ie 0.4 seconds, ends. Figure 2-2 shows an example of a complete QRS complex.

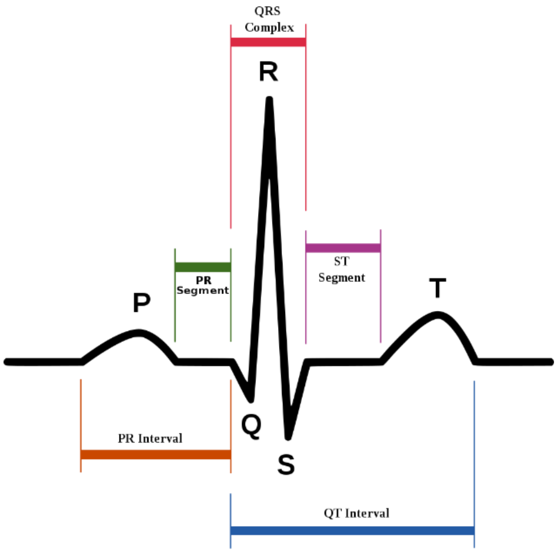


Figure 2:Schematic diagram of a normal human heart rhythm seen by labeling different ECG waves

Covid disease 19

In December 2019, pneumonia due to acute coronavirus syndrome (SARS-CoV-2) was first observed in Wuhan, Hubei Province, China. On February 11, 2020, the World Health Organization (WHO) named the disease caused by infection with (SARS-CoV-2) the 2019 coronavirus (COVID-19). The disease presents with a range of clinical symptoms including fever, dry cough, and fatigue, usually accompanied by pulmonary involvement. SARS-CoV-2 is highly contagious and most people in the community are susceptible to infection. The disease is transmitted through respiratory droplets and direct contact. Since the outbreak of the virus, the Chinese government and scientific community have been quick to identify the cause of the virus and rapidly share the viral gene sequence, and have taken steps to curb the spread.

To date, the virus has caused many casualties around the world and, in addition, has had far-reaching economic and social consequences and increased poverty. Extensive precautions and extensive observance of hygienic protocols such as washing hands regularly, using masks, social distancing, avoiding face-to-face contact, etc. are emphasized to prevent the spread of the virus. Many countries have quarantined their cities to prevent the spread of the virus, as well as enforcing strict laws. The virus quickly affected daily life and business, disrupted global trade, and many countries reduced their production. The industries most affected by the disease are the pharmaceutical industry, the solar energy sector, tourism, information technology and electronics.

Figure 3:corona virus

Data analysis

Data mining is the process of automatically discovering useful information among big data, which includes statistics, machine learning, human-computer interaction, etc. In fact, data mining is the discovery of knowledge in databases that Performs the overall process of converting raw data into useful information. There are several types of data mining, including predictive data mining and descriptive data mining. In predictive data mining, there are two types of variables: objective or dependent variable and explanatory or independent variable. The value of the target variable will be predicted based on the value of the explanatory variables using different techniques. Descriptive data mining also shows the basic relationships in the data and extracts patterns such as clusters and correlations.

In general, data mining techniques are used to find new patterns, predict the results of future observations, and strengthen the information retrieval system, and organizations use data mining and patterns discovered by They can achieve their lofty goals[7].

machine learning

Over the past two decades, machine learning has become one of the most important parts of information technology and can play an important role in any organization.

In general, by computational methods Experience is used to predict or improve performance, say machine learning. Here, experience means prior information that is made available to the learner. This information is usually collected in the form of electronic data and will be available for analysis. The quality and size of this data is critical to the success of the forecasts made, as it will help organizations to ensure that problem-solving results are reliable and prevent irreparable financial damage. Machine learning is composed of efficient and accurate prediction algorithms. In fact, machine learning enables machine learning and the way of this learning will vary based on the type of data and its complexity. Machine learning is divided into two general categories[8]:

- Learning with supervision

Algorithms whose training data are labeled and have a specific purpose are called supervised algorithms. In this type of learning, it is necessary to divide the data set into two categories: training and testing. This segmentation will help to interpret the answers obtained from these algorithms. Training data is a set of data that is considered as an algorithm training term. In this type of learning, there is a set of variables that are considered as input and these input variables affect one or more output variables. In fact, inputs are used to predict output values. So the existence of another category, the test data, is also felt. This type of data is given to the software to evaluate the learning performance performed by the algorithm and according to the results obtained from them, the decision will be made in the organization. The proper performance of the results of this data set is highly dependent on the type of algorithm and the type of data set segmentation.

Supervised learning is used to classify, predict, and fit. To this end, the most important supervised learning algorithms are:

i. Linear Regression

ii. Decision Trees and Random Forests

iii. Logistic Regression

iv. Support Vector Machine

v. K-Nearest Neighbors

vi. Artificial Neural Networks

- Learning without supervision

Unlike supervised learning, in supervised learning, educational data has no label. In unsupervised learning, there are no specific outcomes and goals associated with each input. Instead, the unsupervised learner has previous biases about what aspects of the input structure should be taken at the output, and this will make the gender of the issues in this section different from before. There are many different types of problems in this type of learning that can be referred to clustering and problems related to the diagnosis of anomalies. Issues related to shopping cart analysis and proposing systems can also be considered as part of them. There are different algorithms for unsupervised learning, which are:

i. K-means

ii. Single class backup vector machine

iii. Isolation Forest

iv. DBSCAN

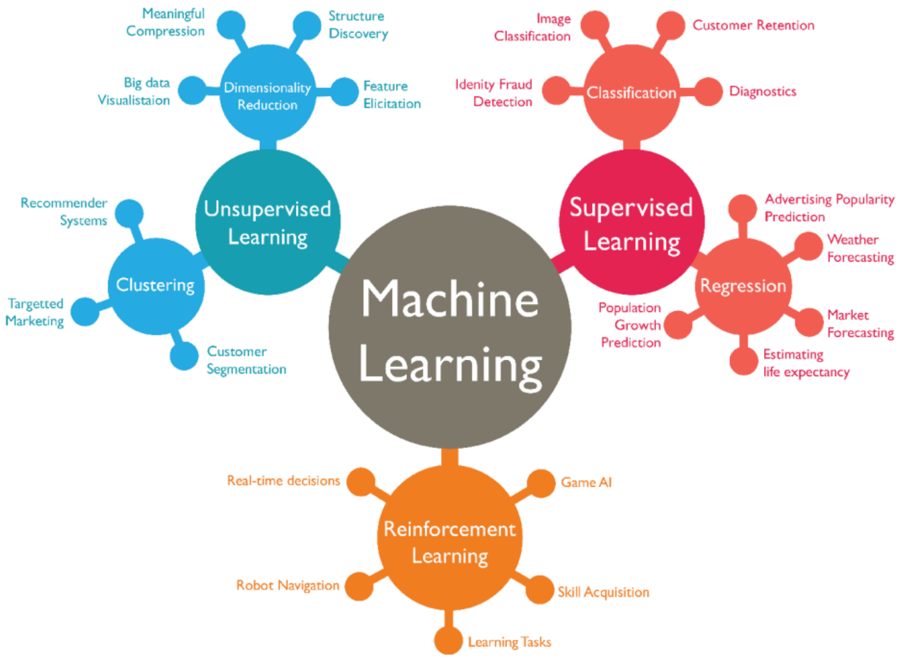


Figure 4: machine learning parts

Chapter 2:

Previous study

Introduction

The widespread prevalence of covid 19 disease, on the one hand, and heart problems and mortality, on the other, have led to a considerable amount of previous research on the disease, its combination, and even the use of emerging scientific achievements such as machine learning.

1. Thakore and et: Redressed QT (QTc) and QRS interims were measured from ECGs performed earlier to mediation or organization of QT drawing out drugs. QTc and QRS interims were assessed as a work of malady seriousness (patients conceded versus released; inpatients conceded to restorative unit vs ICU) and cardiac inclusion (troponin rise >0.03 ng/ml, hoisted B-natriuretic peptide (BNP) or NT pro-BNP >500 pg/ml). Multivariable investigation was utilized to test for centrality. Chances proportions for indicators of infection seriousness and mortality were generated. Baseline QTc of inpatients was drawn out compared to patients released and relative to a control gather of patients with flu. Inpatients with irregular cardiac biomarkers had drawn out QTc and QRS compared to those with typical levels. Discoveries were affirmed with multivariable investigation. QTc prolongation autonomously anticipated mortality. QRS and QTc interims are early markers for COVID-19 malady movement and mortality[10].

2. Hugo De Carvalho and et: Distinguishing proof of at hazard patients and instruments basic cardiac association in COVID-19 remains hazy. Amid hospitalization for COVID-19, tall troponin level has been found to be an free variable related with in-hospital mortality and a more noteworthy hazard of complications. Electrocardiographic (ECG) variations from the norm may be a valuable device to distinguish patients at hazard of destitute prognostic. The point of our study was to evaluate on the off chance that particular ECGs designs can be related with in-hospital mortality in COVID-19 patients displaying to the ED in a European country.we conducted a multicenter think about in three healing centers in France. We included grown-up patients (≥ 18 a long time ancient) who gone to the ED amid the think about period, with ECG performed at ED confirmation and analyzed with COVID-19. Statistic, comorbidities, drug exposures, signs and side effects displayed, and result information were

extricated from electronic therapeutic records employing a standardized information collection shape.The relationship between ECG anomalies and in-hospital mortality was evaluated utilizing univariate and multivariable calculated relapse analyses. An ECG was performed on 275 patients who displayed to the ED. Most of the ECGs were in typical sinus beat (87%), and 26 (10%) patients had atrial fibrillation/flutter on ECG at ED confirmation. Repolarization variations from the norm spoken to the foremost common discoveries detailed within the populace (40%), with negative T waves speaking to 21% of all variations from the norm. We found that unusual hub, and cleared out bundle department piece were essentially related with in-hospital mortality[11].

3. Luca Bergamaschi and etc: We assessed 269 successive patients conceded to our center with affirmed SARS-CoV-2 disease. ECGs accessible at affirmation and after 1 week from hospitalization were evaluated. We assessed the relationship between ECGs discoveries and major antagonistic occasions (MAE) as the composite of intra-hospital all-cause mortality or require for obtrusive mechanical ventilation. Unusual ECGs were characterized in case essential ST-T portion modifications, cleared out ventricular hypertrophy, tachy or Brady arrhythmias and any modern AV, bundle squares or critical morphology changes (e.g. modern Q neurotic waves) were show.Anomalous ECG at confirmation and hoisted pattern troponin values were more common in patients who created MAE.Concerning ECGs recorded after 7 days, unusual discoveries were detailed in 53.5% of patients and they were more visit in those with MAE. Among anomalous ECGs, ischemic changes and cleared out ventricular hypertrophy were essentially related with the next MAE rate. The multivariable examination appeared that the nearness of anomalous ECG at 7 days of hospitalization was an free indicator of MAE. Moreover, patients with unusual ECG at 7 days more frequently required exchange to the seriously care unit or renal substitution treatment[12].

4. Esmaeil Mehraeen and etc: We conducted a efficient look in PubMed, Embase, and Scopus databases. The group recognized 20 articles related to this theme. We separated them into articles talking about drug-induced and non-drug initiated changes. Considers detailed an expanded chance

of QTc interim prolongations impacted by distinctive treatments based on chloroquine, hydroxychloroquine, and azithromycin.In spite of the fact that these drugs expanded dangers of serious QTc prolongations, they actuated no arrhythmia-related passings. Within the non-drug-induced gather, ST-T anomalies, outstandingly ST elevation, accounted for the foremost watched ECG finding within the patients with COVID-19, but their connection with myocardial injuries was under dispute. This orderly audit proposes that distinguishing ECG designs that may well be related to COVID-19 is crucial. Given that doctors don't recognize these designs, they might wrongly hazard the lives of their patients. Moreover, critical drug-induced ECG changes give mindfulness to the health-care laborers on the dangers of conceivable therapies.[13]

Chapter 3:
Modeling and solution method

Covid data set 19

Introducing the data set
The data set under study contains 688 features, which we will briefly introduce in the following.

we obtained the set of features from 12-lead ECG signals of 10 seconds. Features were obtained using two commercial softwares (Bravo and Glasgow). Signals were acquired at Monzino Hospital.

In this data set, we examined the data of patients with Covid-19 and classified the patients. In this design, the target is initially classified as binary (dead and alive) and in the research stages, these changes are divided into two classes and the classes are divided according to the data related to the properties. This division will be shown in the figure below.

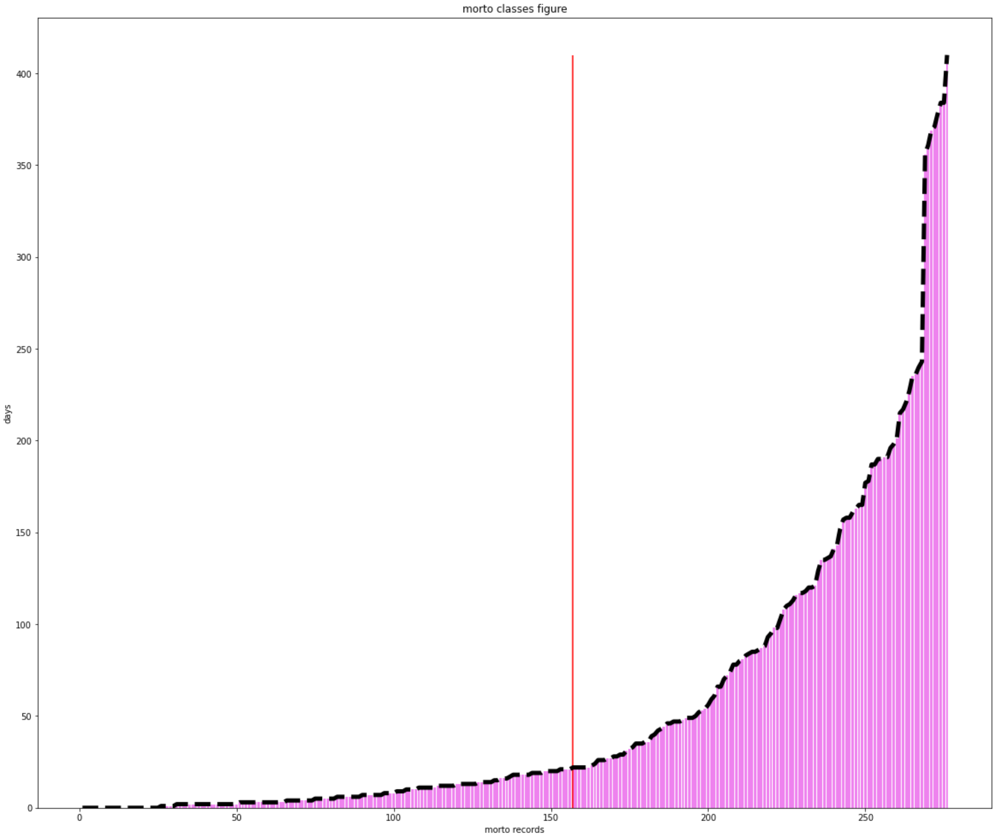


Figure 5: Classification of patients leading to death over time

The process of reviewing data from beginning to end

As mentioned, in this study, data set related to Covid 19 and patient mortality prediction have been used. There must be a process to reach the goal. This process is schematically shown in Figure 2. Each part of this process will be described in detail.

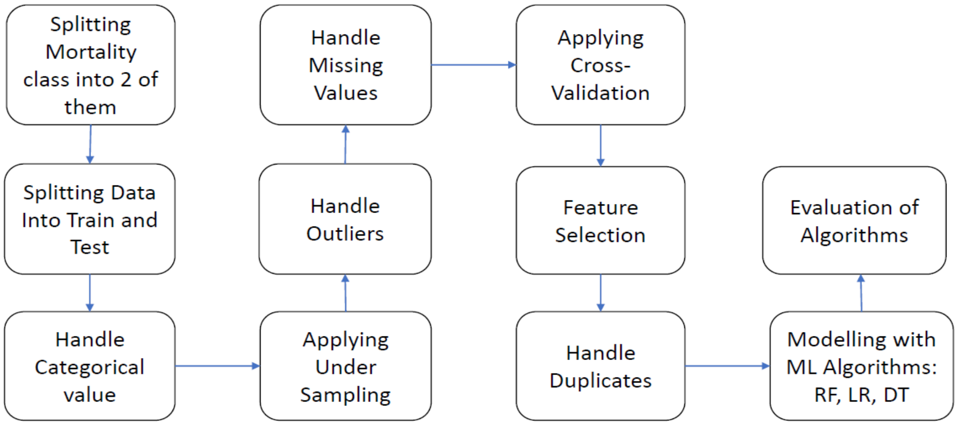


Figure 6:The process of reviewing a data set from the beginning to achieving the goal

At the beginning of this process, the data of the dead were divided into two classes. The main data set is divided into two categories: training data and test data with a ratio of 80 to 20. The main activity should be done on educational data. Because ultimately the use of machine learning algorithms depends on the correct use of this data. Then, preprocessing operations were performed on these data. First, non-numerical data were examined, but with the performed studies, this data set did not have any non-numerical data. Then, in the next step, the nearest neighbor method was used to deal with the lost data. This method uses other data that is safe, and in other features, those that have values close to the data, use its value and use it in the lost data locus. In this way, all the lost data was correctly replaced with complete data. In the next step, when it came to outliers, Tukey statistical method was used to pursue two goals. First identify these data and then decide on the appropriate treatment according to the following formula and box diagrams. Outdated data is so-called unreliable data in the data set and their existence may negatively affect the final performance obtained by machine learning algorithms.

The first quarter in the box diagram = Q1

The third quarter in the box diagram = Q3

Upper Bound = Q3 + 1.5 * (Q3 – Q1)

Lower Bound = Q1 – 1.5 * (Q3 – Q1)

The box plots below show the data scatter for each feature and the different target feature classes. If the data is higher or lower than the lower limit, it is considered out of date data. Each diagram in each image belongs to a column of the data set. Different columns have different output data.

The following figure belongs to 3 classes 0, 1 and 2. For example in P_Area_V1_2 the data density is between -2.5 and 0.5.

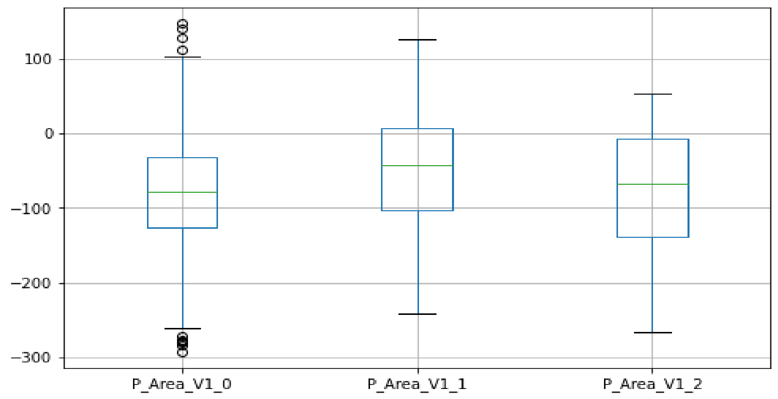


Figure 7:P_Area_V1 box plot

In P_Area_V4_1 the data is between -1.75 and 1.75 and there is a drop of data above the upper limit.

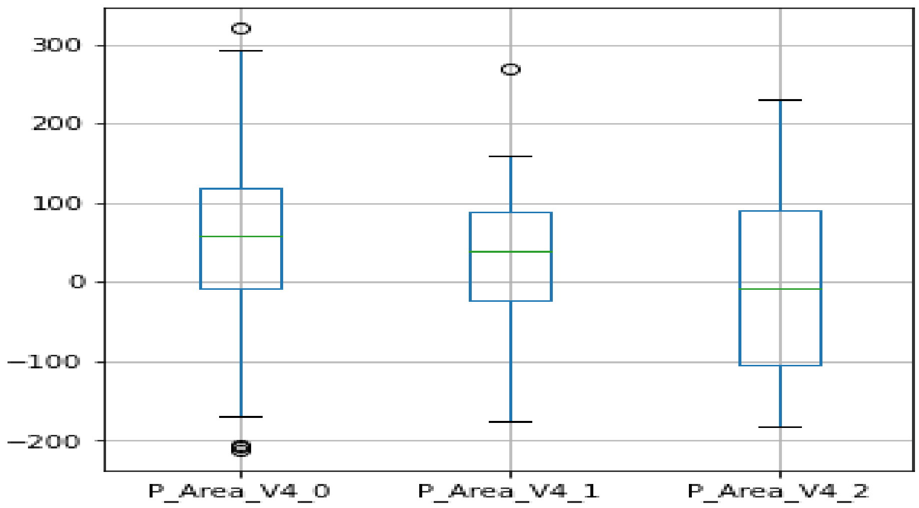


Figure 8: P_Area_V4 box plot

In the P_Morph_V6 feature, the box diagram is shown above because the numeric value 1 alone is in the training dataset. Naturally, the data density will be exactly one because there is no other numeric value.

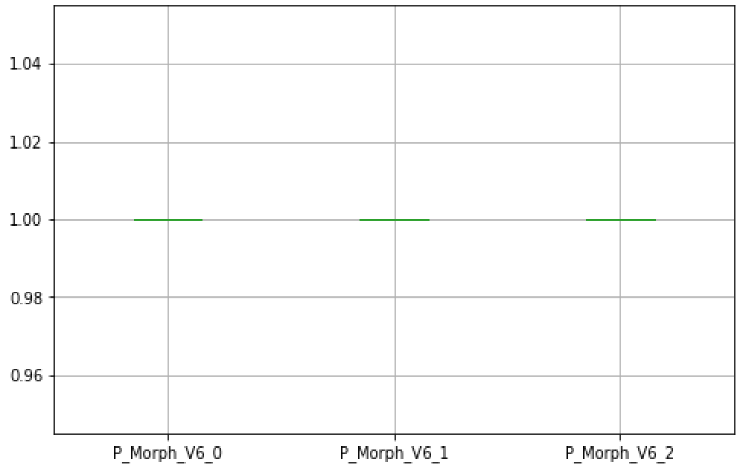


Figure 9:P_Morph_V6 box plot

In QRS_Pseudo_Vector_7 / 8_Spatial_Velocity_0 there are two outliers above the upper limit.

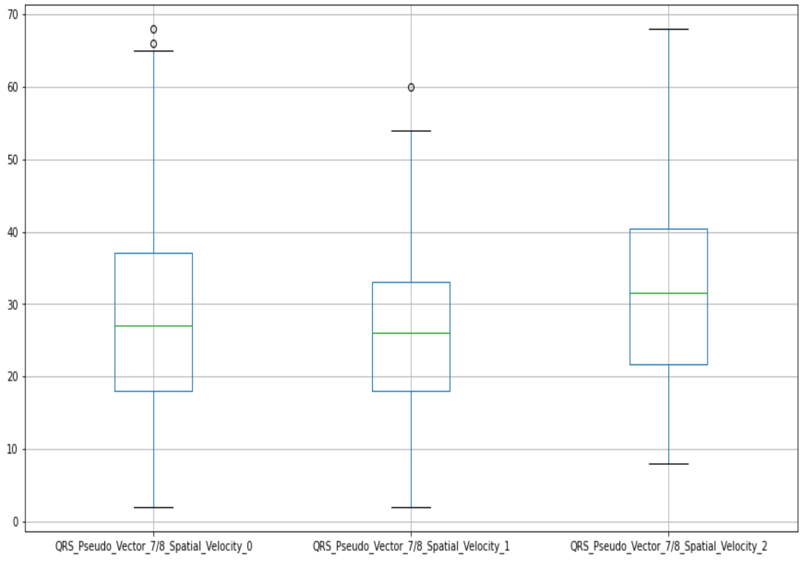


Figure 10: QRS_Pseudo_Vector_7 / 8_Spatial_Velocity

In R_Dur_II_0, the two points 21 and 65 are the upper and lower limits, and in this pattern, there is no outbound data, but in the rest of the classes related to the deceased, there is outbound data.

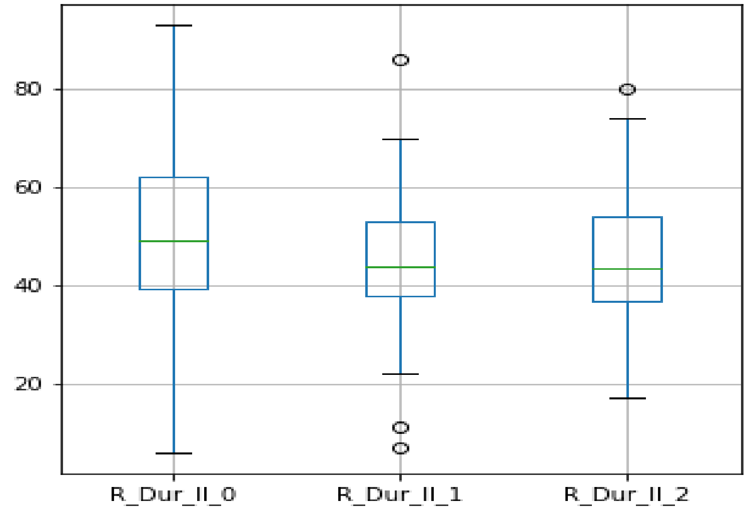


Figure 11: R_Dur_II

In the R_Notch_V2 attribute, the box diagram is shown above because the numeric value zero alone is in the training dataset. Naturally, the data density will be exactly one because there is no other numeric value.

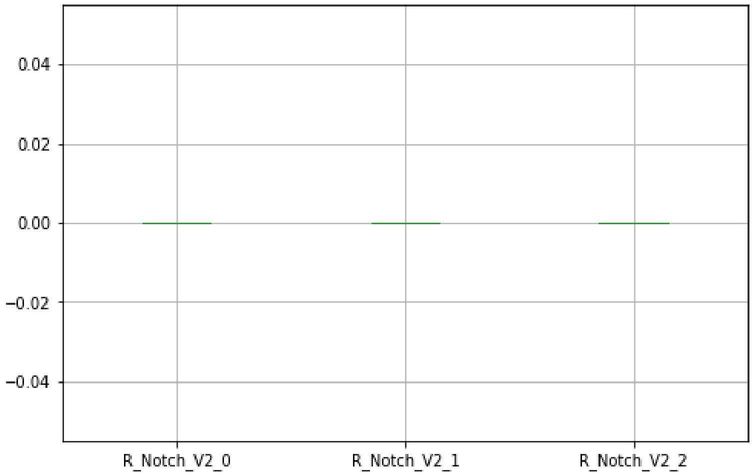


Figure 12: R_Notch

STamp_2.0_Values_2 shows two outbound data, one above the upper limit and the other below the lower limit. In other words, the output data is seen for all classes.

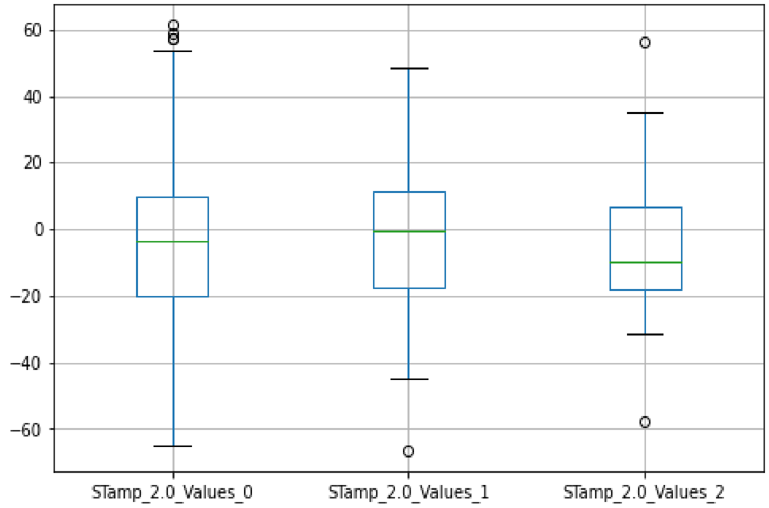


Figure 13: STamp_2.0_Values

T_plus_Amp_V3_0 shows some junk data that is above the limit. These data belong to patients who survived the infection.

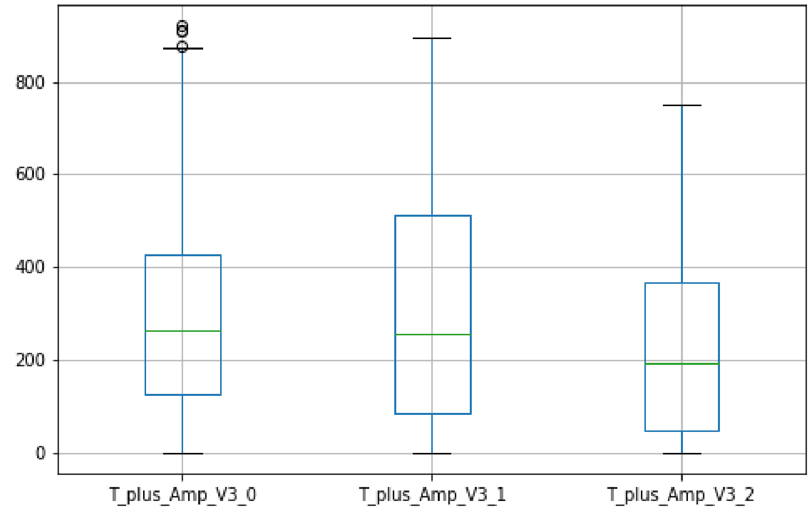


Figure 14: T_plus_Amp

In Age_2 there are three outbound data, one of which is higher than the upper limit and the other two are lower than the lower limit. Part of this data belongs to the age of living after the disease and the rest belong to the deaths after the disease.

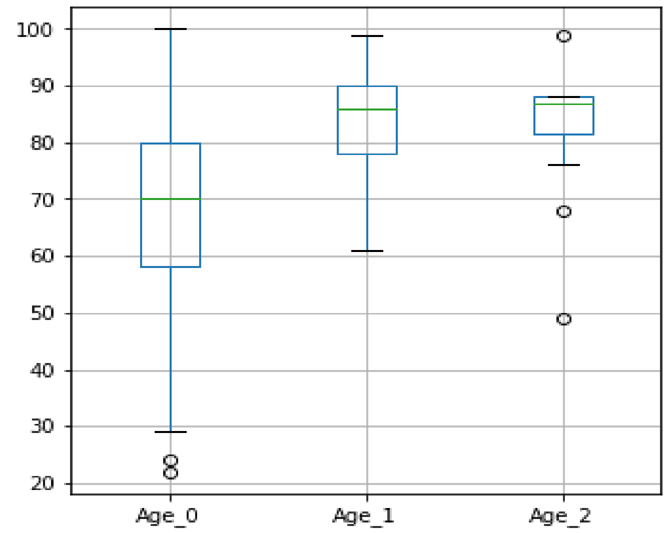


Figure 15: Age

In PP_3.0_values_1, the data density is between 100 and 1150 and there is no data output.

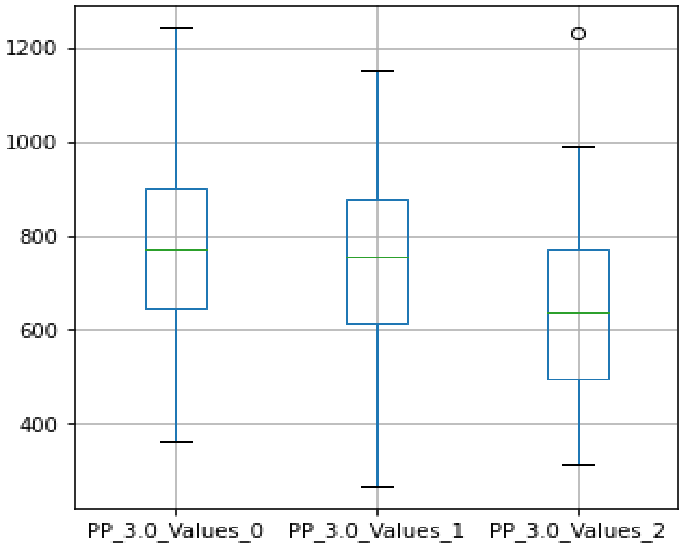


Figure 16: PP_3.0_values

PR_Interval_1 shows two outbound data, one above the upper limit and the other below the lower limit.

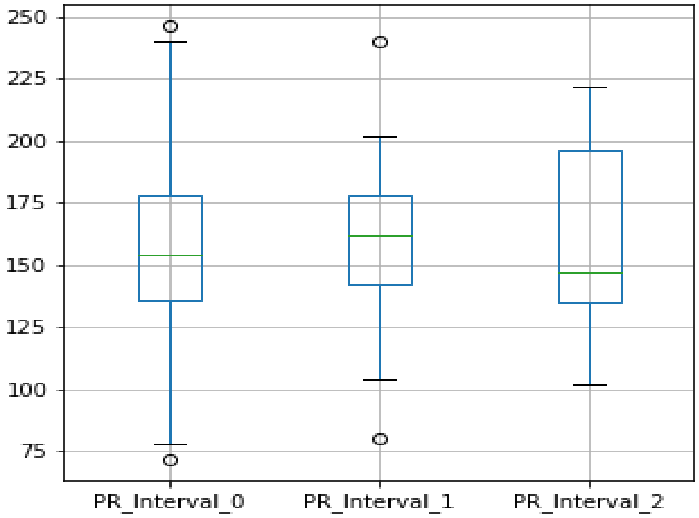


Figure 17: PR_Interval

In S_amp_v6_2 the data is between 0 and -350. And there are no data outside the classes of the deceased and living people.

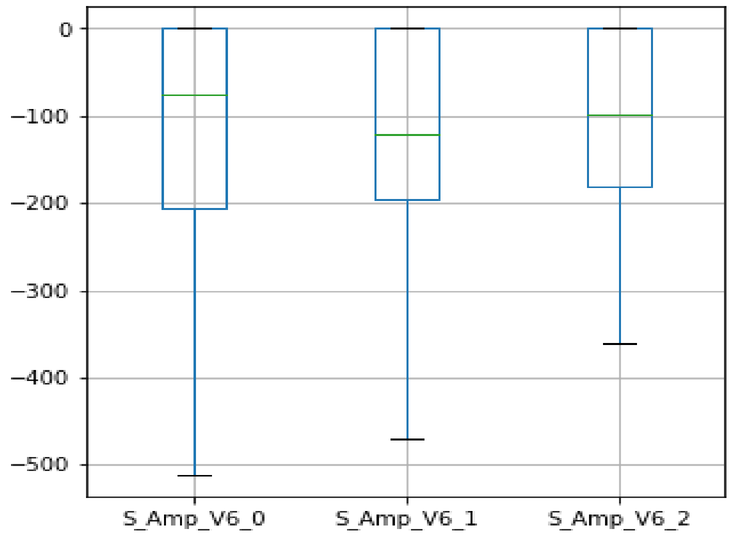


Figure 18: S_amp_v6

After carefully reviewing and dealing with the existing outdated data in accordance with the objectives of the problem, it was time to review the duplicate data. This type of data is called duplicate data when two or more rows are repeated in the same data set. This particular type of data causes the expected performance of the machine learning algorithms used to be inaccurate and the decision-making organization will eventually face many problems. After careful examination of the dataset, no duplicate data were found.

After these steps, one of the most important steps must be done. This step will be called "Feature Selection". Choosing the right features is important because it eliminates irrelevant and additional features in the first place, as well as reducing the complexity of the data set to enter as input to machine learning algorithms. In other words, this action will make the expected final performance optimal and quality.

There are different ways to choose the features:

1. Filter method

This method uses statistical techniques such as the Khuk Dual Test. The points earned according to the P-Value will help us choose the best feature.

1. Wrapper method

This method is used when the complexity of the data used is usually high. There are various methods for it, the most prominent of which will be Recursive Feature Elimination, in which the research is used to select the features to investigate the Support Vector Machine algorithm. And due to the structure of this algorithm, it has used it and reduces the results of the final features to 4 features. The results are in the table below. The choice of feature uses methods that select a subset of the most relevant features (columns) in a dataset. The lower the features and the higher the characteristics related to the target feature, the machine learning algorithms will be better.

RFE is a feature selection algorithm. In fact, RFE selects the properties in the training dataset that have the greatest impact on predicting the target variable.

RFE has two important meta-parameters:

1. Number of attributes to be selected.

2. Algorithm used to help select features.

Technically, RFE is a packaging-style feature selection algorithm that also uses filter-based feature selection inside.

To achieve a subset of features, the RFE fits the algorithm used, ranks the features by importance, discard the least important features, and re-fits the model. This works as long as a certain number of features remain. Is repeated.

Table 1: features selected by the Wrapper method

number	Features name
1	P_Area_V1
2	P_Area_V4
3	P_Morph_V6
4	QRS_Pseudo_Vector_7/8_Spatial_Velocity
5	R_Dur_II
6	R_Notch_V2
7	STamp_2.0_Values
8	T_plus_Amp_V3
9	Sex
10	Age
11	PP_3.0_Values
12	PR_Interval
13	S_Amp_V6

3. Embedded method

This method also uses machine learning algorithms to select features. Of course, in this method, algorithms are used that are based on the term tree. Algorithms such as decision tree, random forest, etc. are among these methods.

After selecting the appropriate properties, the following figure shows the relationships between the main variables of the data set under study.

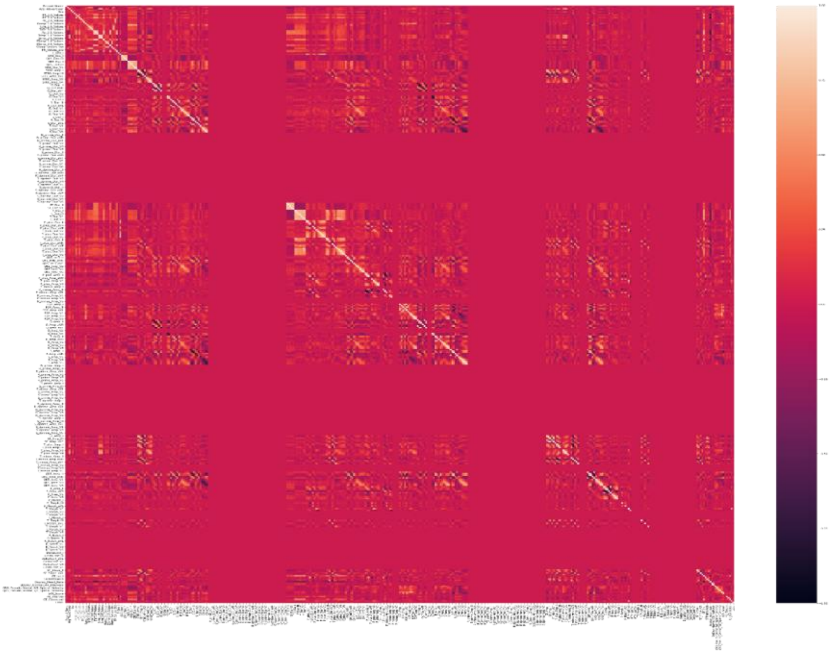


Figure 19: correlation matrix before feature elimination

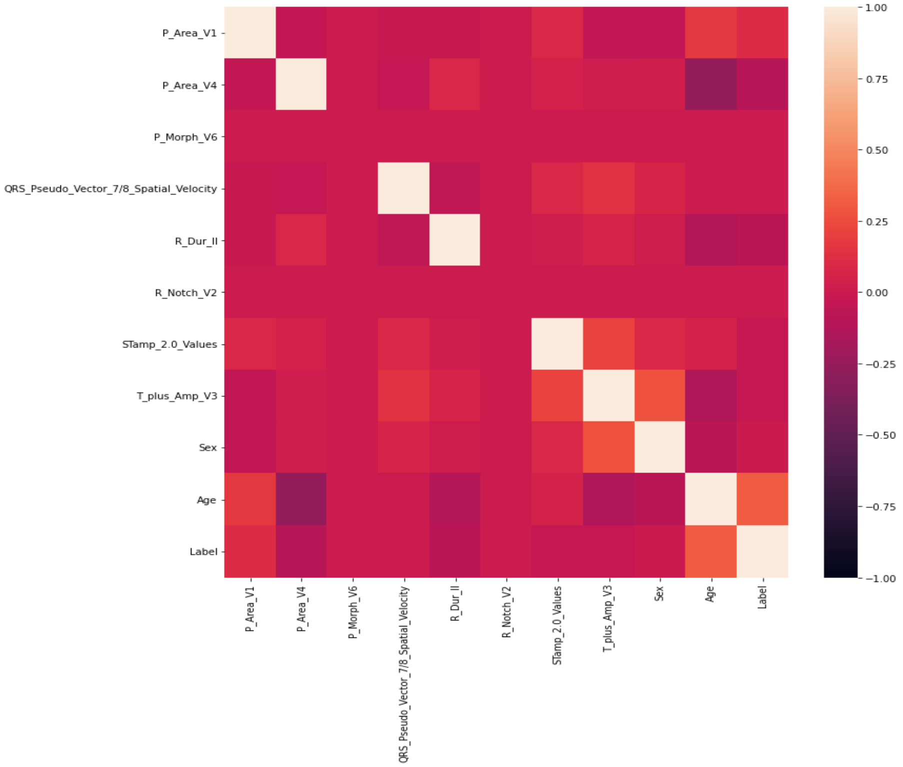


Figure 20: correlation matrix after feature elimination

After the feature selection step, the Cross-Validation step must be performed. In this section, the StratifiedKFold method is used, which is first divided into 3 parts of the training data training results as follows:

Table 2:Results obtained from the division of training data by Cross-Validation method

methods	Earned values
The accuracy of each division	[0.86315789 0.8556338 0.87323944]
Accuracies average	0.8640103780578207
standard deviation	0.007212703203798695
Accuracy assurance interval	(0.8567976748540219, 0.8712230812616194)

One of these divisions will then be used and the educational data will be divided into validation data.

Due to the classification and division of data into educational and validation data, data distribution histogram will be as follows in each classification.

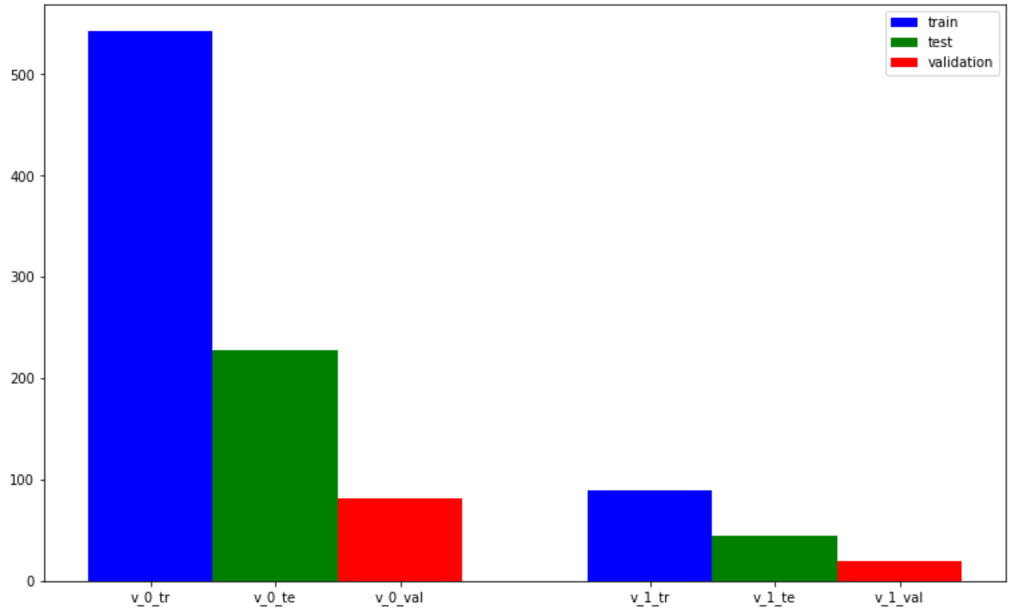


Figure 21: Histogram of data distribution according to their division into training, test and validation

Given the difference in the number of data in each class and the superiority of the class of living people to those who died after coronary heart disease and their imbalance, a solution must be considered to balance. For this purpose, the under sampling technique has been used. In this method, due to the 5-fold superiority of the living over the dead, the data were divided into 5 main subsections. Also, due to the fact that the deceased were in two classes, the data became unbalanced again. For this reason, 2 sections are added to each of the 5 subsections. Finally, according to this technique, the data will be ready for training by special machine learning algorithms.

Machine learning methods used in research problem modeling

In a data analysis project, fuzzy modeling is used in which machine learning algorithms are used to identify specific patterns in the data set. A machine learning model is a set of algorithms that analyzes large volumes of data to find patterns or predictions. Machine learning models are artificial intelligence math engines that feed on data. In each data analysis project, according to the purpose and nature of the problem, special models are used to answer the problem. In this research, algorithms and methods of random forest (XGBoost) and decision tree (Decision Tree) have been used to achieve the goal. In the following, each of them is described in detail in separate sections.

Random forest algorithm

Stochastic forest algorithms are part of a group of algorithms that include tree-based algorithms and ensemble learning. In hybrid learning, a large number of basic models are combined to achieve an optimal prediction. Random forest algorithms use a large number of decision trees for training. Eventually the prediction of all these trees is combined for the final prediction. In this method, for final decision making, fashion is used in classification models, and in regression models, the mean is used.

Decision Tree Algorithm

Decision Tree Algorithm is a supervised learning algorithm used in both classification and regression problems.

One of the attractive features of this algorithm is the graphic or model tree diagram that this tree pattern is composed of nodes, branches and leaves. Nodes represent properties, branches represent decisions, and leaf nodes represent output.

In this algorithm, the sample is divided into two or more homogeneous sets based on the divider.

The decision tree uses different algorithms to decide whether to split a node into two or more sub-nodes. Node purity increases according to the target variable. The choice of algorithm is also based on the type of target variables. The decision tree divides the nodes by all available variables and then selects the divider that gives the most homogeneous sub-nodes.

Impurities in the decision tree are very important because it measures the degree of homogeneity in a data sample, and if the sample is homogeneous, it is a sample of the same class. There are several criteria for measuring impurities in a sample, two of the most important of which are:

Entropy index and Gini index

entropy
Entropy describes the required sample. If the sample is homogeneous, it means that all elements are the same and the entropy is 0; otherwise, if the sample is evenly divided, the entropy is a maximum of 1.

Entropy is written mathematically as follows:

$$Entropy = -\sum_{i=1}^{n} p_i * \log(p_i)$$

Gini index
The Gini index measures inequality in the sample and its value is between 0 and 1. If its value is zero, it means that the sample is completely homogeneous and all elements are the same, and if its value is one, it means that there is a maximum inequality between the elements.

The Gini index is the sum of the probabilities of each class and is displayed as follows:

$$Gini = 1 - \sum_{i=1}^{n} {p_i}^2$$

Data algorithms are used to produce decision -making tree:

1.Cart

2.ID3

3.chaid

4.ID 4.5

Of these algorithms, Cart and ID3 are most commonly used.

Intense Gradient Amplification Algorithm (XGBOOST)

The algorithm was first started as a research project by Tiangi Chen as part of a distributed deep machine learning team, and after winning the Higgs Machine Learning Challenge, it became known as part of the race cycle. This algorithm is one of the tree-related algorithms that is used to reinforce the answers obtained from machine learning models.

Algorithm N Educational data $\{(x_i, y_i)\}_{i=1}^{N}$ takes a derivative cost function L (y, F (x)) and an alpha learning rate as input and follows the steps to build the model Final pays:

Step One) We have a weak model that minimizes the cost function:

$$\hat{f}(x) = argmin \sum_{i=1}^{N} L(y_i, \theta)$$

Step 2) In this step, assuming m = 1,…, M:

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)} .$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)} .$$

Final Step) Given the above steps the final model will be as follows:

$$\hat{f}(x) = \sum_{m=0}^{M} \widehat{f_m}(x)$$

Logistic Regression

Logistic regression is a statistical method used when the dependent variable is binary or dual. This technique is used to predict probabilities for classification problems. Logistic regression describes the relationship between a dependent variable and a set of independent variables[9].

Hypothesis function for logistic regression

The output of the hypothesis function is the estimated probability. This function is used to measure how close the predicted value is to the actual value. The function of the hypothesis used in logistic regression is as follows:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

And the Sigmid function is as follows:

$$h_\theta(x) = sigmoid\ (z)$$

If Z tends to be infinitely positive, the predicted value is 1, and if Z tends to be infinitely negative, the predicted value is 0.

Cost function

The accuracy of the hypothesis function can be calculated with the cost function. The cost function for logistic regression is as follows:

$$h_\theta(x), y) = -\log(h_\theta(x)) \quad if\ \ y = 1$$

-log($h_\theta(x)$) $if\ y = 0$

The tips in this function are as follows:

If y = 0 and h_θ (x) tends to zero then cost = 0.

If y = 0 and h_θ (x) to one mile then cost tends to infinity.

If y = 1 and h_θ (x) tends to zero then cost tends to infinity.

If y = 1 and h_θ (x) go to the same desire then cost = 0.

The cost function can also be summarized as follows:

$$\text{cost} h_\theta(x), y) = \; -(1-y) log h_\theta(x)) - y \log(\, h_\theta(x))$$

J(θ) = 1/m$\sum_{i=1}^{m}$ Cost$\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) = -\frac{1}{m}[\sum_{i=1}^{m} \left(1 - y^{(i)}\right)\log\left(1 - h_\theta\left(x^{(i)}\right)\right) + y^{(i)} \log(\, h_\theta\left(x^{(i)}\right))]$

Now J (θ) must be converge to the minimum universal, which is used to do this:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)} - y^{(i)})\, x^{(i)}$$

Chapter Summary

As stated, in every problem and with every goal, the challenges created need to be solved and appropriate methods must be used. It was also stated that Covid 19 disease is a very dangerous disease that has caused many deaths around the world. For this reason, several researches have been done in this field, each of which, according to their goals and challenges, have used different methods to solve them. In this chapter, with the aim of classifying patients with Covid 19 and selecting appropriate machine learning algorithms, various analyzes were performed on the data set. First, the data were analyzed and pre-processed, and after clearing and selecting the appropriate features, the problem was balanced between the data. Finally, the training data were prepared by the machine learning algorithms used in this chapter.

In the next chapter, while introducing the methods and techniques of evaluation of the mentioned algorithms, the results of the used algorithms will be described according to them.

Chapter4:

Evaluation of results

Introduction

In each algorithm, according to the goals it pursues, special techniques are used to validate the model.

In this section, various methods are used to measure the performance of each algorithm. All the methods used in this research will be described in the following sections.

Introducing the methods of evaluating the results used in the research

Confusion matrix

The clutter matrix is a way to represent the performance of classification algorithms. In this matrix, the columns represent the prediction items and the rows the actual instances of the classes. This matrix is also called the error matrix learning machine. The main diameter of the matrix shows the correct predictions of the algorithm and the sub-diameter of the prediction errors. Using this matrix, the accuracy of the algorithm can be evaluated. This matrix consists of several sections that describe the values inside the matrix

1. True Positive Predictable Values
2. True Negative and True Predicted Values
3. False Positive Real Values
4. False Negative Real Values

Precision and Recall

Precision and recall are two criteria used to assess the sensitivity and recognizability of models and are commonly used when data are unbalanced.

In information retrieval, precision is a measure of the relevance of results and is known as a positive predictive value, while recall is a measure that is returned from the number of related actual results and is known as sensitivity in binary classification.

precision The number of real positive results is divided by the number of all positive results and its formula is as follows:

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

recall The number of real positive results is divided by the number of all samples that should have been positive and its formula is as follows:

$$Recall = \frac{true\ positive}{true\ positive + false\ negetive}$$

A system with high recall but low precision returns many results, but most of its predicted labels are incorrect compared to training labels.

In contrast, a system with high precision but low recall returns little results, but most of its predicted labels are correct compared to training labels.

An ideal system is one that has high precision and high recall. Such a system will have many results and all the results are properly labeled.

F-Score

F-score or F-measure is a measure of the accuracy of the model. F1-score can also be considered as the harmonic mean of precision and recall, in which the F1-score reaches its best value at 1 and its worst value at 0. The F1-score formula is as follows:

$$F1 = 2 \times \frac{precision\ \times recall}{precision + recall}$$

Accuracy score

Accuracy is a measure for classification models that measures the number of correct predictions as a percentage of the total number of predictions made. When data is unbalanced, accuracy is not a good benchmark for use. The formula for calculating the accuracy of the model is as follows:

$$Accuracy = \frac{true\ positive + true\ negetive}{true\ positive + true\ positive + true\ negetive + false\ negetive}$$

The results obtained according to machine learning methods

The values obtained for the disruption matrix for testing and validation data are as follows:

Table 3: Results from the confusion matrix associated with validation data

TP = 227	FP = 7
FN = 31	TN = 19

As shown in Table 1, the values from the validation data, for example, that are positive and correctly predicted, include 227 data. Other values are also described in the table. Table 2 also shows the confusion matrices associated with the test data.

Table 4: Results obtained from the confusion matrix associated with the test data

TP = 118	FP = 6
FN = 9	TN = 18

In this table, for example, it can be seen that negative values are erroneously predicted, which is also considered as an error in the test toast of 9 toasts.

The following table shows the values of the results for binary classes according to the evaluation criteria of the algorithms. In this section, the target variable is considered as zero and one.

Table 5: Results obtained from binary class data according to evaluation criteria

Method	Accuracy Score	Sensitivity	Specificity	F1-Score
Class0	0.9006622	0.9	0.81	0.86
Class1	0.9006622	0.8971	0.9007	0.8983

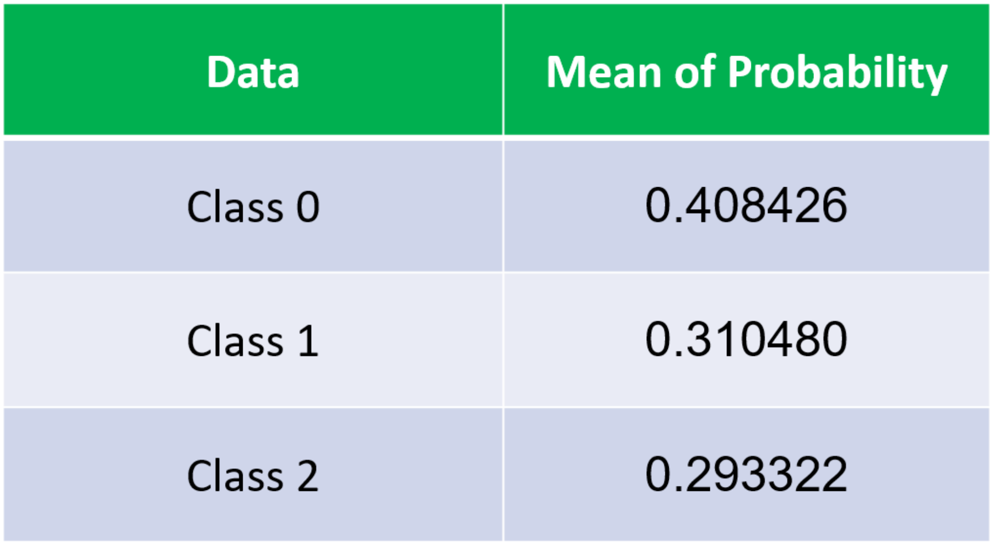
The table below shows the values listed for the case where the target variable is classified into three classes. In fact, these values are displayed separately for each class.

Table 6: Results obtained from 3-class data according to evaluation criteria

Method	Accuracy Score	Sensitivity	Specificity	F1-Score
Class0	0.737402	0.631776	0.597293	0.609810
Class1	0.615376	0.482764	0.396785	0.396783
Class2	0.642832	0.431657	0.398147	0.378421

In Table 7, depending on the classes, the average probability of each class is based on machine learning algorithms. This means that, for example, the probability of predicting the algorithms is zero at an average of 40 %. While this value is 31 and 29 percent, respectively for the Class 1 and 2, which shows the different classes of the deceased. Obviously, the sum of these possibilities will be equal to 1.

Table 7: Average probability of occurrence of different classes of data set in prediction by algorithm



Data	Mean of Probability
Class 0	0.408426
Class 1	0.310480
Class 2	0.293322

Season Summary

As stated, in each algorithm according to the goals it pursues, specific techniques are used to validate the model. This chapter first introduced validation methods tailored to the data set used in this research. After that, the results of the machine learning algorithms applied to the dataset were examined in accordance with different classes. Finally, by each class, the average probability of each class was obtained according to the applied machine learning methods.

References

1 Hong He , Yonghong Tan, Jianfeng Xing, (2019).
"Unsupervised classification of 12-lead ECG signals using wavelet tensor decomposition and two-dimensional Gaussian spectral clustering"

2 Esmaeil Mehraeen a, Seyed Ahmad Seyed Alinaghi,etc.
"A systematic review of ECG findings in patients with COVID-19"

3 Saman Parvaneh a,⁎, Jonathan Rubin a, Saeed Babaeizadeh b, Minnan Xu-Wilson(2019)
"Cardiac arrhythmia detection using deep learning: A review"

4 Keith, L., A. Moore, and A. Agur, "*Clinically oriented anatomy".*

5 Starr, C., C. Evers, and L. Starr, "*Biology today and tomorrow with physiology".*

6 Serhani, M.A., et al., "*ECG monitoring systems: Review, architecture, processes, and key challenges.*"

7 Brown M. Data Mining For Dummies. 2nd ed.

8 Wang H., Zheng H. (2013) "Model Validation, Machine Learning. In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY."

9 C., pampel, Fred. "Logistic Regression: A Primer (Quantitative Applications in the Social Sciences)"

10 Avni Thakorea,*, James Nguyen,etc. "Electrocardiographic manifestations of COVID-19: Effect on cardiacactivation and repolarization"

11 Hugo De Carvalho1, Lucas Leonard-Pons,etc. "Electrocardiographic abnormalities in COVID-19 patients visiting the emergency department: a multicenter retrospective study"

12 Luca Bergamaschi MD1 | Emanuela Concetta D'Angelo MD1,etc. "The value of ECG changes in risk stratification of COVID-19 patients"

13 Esmaeil Mehraeen[a], Seyed Ahmad Seyed Alinaghi,etc. “A systematic review of ECG findings in patients with COVID-19”