International Journal of Advance Research and Innovation
Vol. 10(3), Jul-Sept 2022, pp. 67-73
Doi: 10.51976/ijari.1032209
www.gla.ac.in/journals/ijari

Received: 15 May 2022 | Revised Submission: 15 Jul 2022 | Accepted: 05 Sept 2022 | Available Online: 15 Sept 2022

A Study on Data Cleaning using Visualization by Prediction and Health Monitoring

Manasvi Dev, Mansi Singh**, Vidhi Shah***, Amit Hatekar**** and Manoj Kevadia******

ABSTRACT

This paper presents an application through which any medical dataset could be cleaned, visualized and used for prediction. A health checkup system for a patient's vitals has been developed using the Internet of Things. Here we discuss in detail about user friendly reports generated through various graphs. Compared with several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 89.99% with a convergence speed which is the Random Forest prediction algorithm.

Keywords: *Medical datasets; Data cleaning; Prediction; Visualization; Internet of things.*

1.0 Introduction

Irrespective of the way data is collected, there will always be some sort of error and inconsistency. Due to this data cleaning is one of the most important processes done in data mining. The failure in dealing with this issue can result in inaccurate analyses and poor decision making especially if the large quantities of data comes into the picture. In today's world many different sectors highly rely on data analysis and prediction to ensure that they can succeed. Due to this data cleaning becomes a vital prerequisite, as it helps to remove any form of inconsistent or distorted entries making the data more refined and structured.

1.1 Need for data cleaning

Healthcare requires a large amount of dataset to understand the patterns and for diagnosis. Low data quality becomes a major issue in the medical domain as there is a high need for effective decision making and poor data quality can have an adverse effect on people's lives. Dealing with duplicated, incomplete, and inaccurate data is the most important issue faced by the medical sector while handling data. According to many studies, duplicate records make up 5 to 10 % of the dataset. Lastly, if there is inaccurate data present then it makes analysis more tedious and lessens the leverage for better insights or outcomes. Also, inaccurate data of the patients' contact details can led to tedious communication with them and can cause further complications.

1.2 Problem Statement

The start of an early treatment can drastically reduce the odds of a patient's degrading health. In this project we are mainly going to focus on two parts, the first which is the most important part is the data cleaning and processing of various diseases datasets while the second part deals with visualization of predicted outcomes. The outcomes are achieved by passing the cleaned data under various algorithms to obtain the most accurate result.

2.0 Related Work

There are various existing studies carried out by many authors on data cleaning and visualization. The aim of this approach was to generate an abstract clean instance which is the perfect approximation of all feasible concrete clean instances. The ability to research and present data in a clear manner is critical to the success of public health surveillance. Health researchers need useful and intelligent tools to assist their work. The data mining technique can be used to extract production rules from healthcare data and

**Corresponding author: Thadomal Shahani Engineering College, Mumbai, Maharashtra, India*
(E-mail: manasvidev.md@gmail.com)
,*Thadomal Shahani Engineering College, Mumbai, Maharashtra, India*
****, *****Department of Electronics and Telecommunication, Thadomal Shahani Engineering College, Mumbai, Maharashtra, India*

clinical diagnosis. Researchers have extracted diagnostic rules from survival data using data mining techniques. The rules extracted using the data mining techniques are similar to those generated manually by experts. Hence, the results of data mining can be easily validated by domain experts. Moreover, the data mining techniques can also be applied in medical databases, which aim to find new medical knowledge. Future disease prediction is very crucial and important for patients with chronic diseases. Many disease prediction models are proposed in the recent past.

In our proposed work, data collection models are developed taking physiological parameters and hidden symptoms of the diseases of the patients. The area of health in recent years has been rapidly integrating technology within the monitoring, diagnosis, and treatment of patients remotely and in place. Thus achieving to enhance the standard of lifetime of patients and greater traceability of data from them. Most studies reviewed point to chronic disease monitoring in particular as in 12, 13, 14 which are responsible for the first one monitoring sleep through a gyroscope Smartphone, the second remote monitoring of vital signs, and the third of a telemedical ECG system of a patient.

3.0 Methodology

In this section, we will learn about the detailed working of an application with the block diagram and the flow of the application. The cleaning of data that the user uploads to the program are the first phase. In this paper, a benchmark medical dataset has been used. The data is cleaned before predictive modeling. The CSV file or the xlsx file is taken as input and the cleaning algorithm is performed by using the Report Profiling module of Python. Once this is done, a Profile report of the data is generated with a detailed explanation and information. The missing value is replaced with 0 internally with the help of the module.

The accuracy of this cleansed data is 98 percent The cleaned data is then downloaded and uploaded to the visualization page where the user can check the graphical format of the data and can change the parameters according to the requirements. In visualization, a user can choose from the different graphs listed in the application. After this, the cleaned data can be uploaded to the prediction page where the prediction of a dataset is done and the confusion matrix is displayed. It has an accuracy of 98% and it gives information on whether or not a person is prone to a particular disease or not by checking different parameters of data.

The application also has a hardware part that takes real-time data from patients and performs predictive modeling on it. It predicts the disease by taking data from users and comparing it with the predictive analysis of the dataset. The output is then displayed on the web page.This was about how an application worked. It almost has an accuracy of 99% and can be used for the research of medical hospitals for insights from the data and help for making a correct decision.

4.0 Data Cleaning

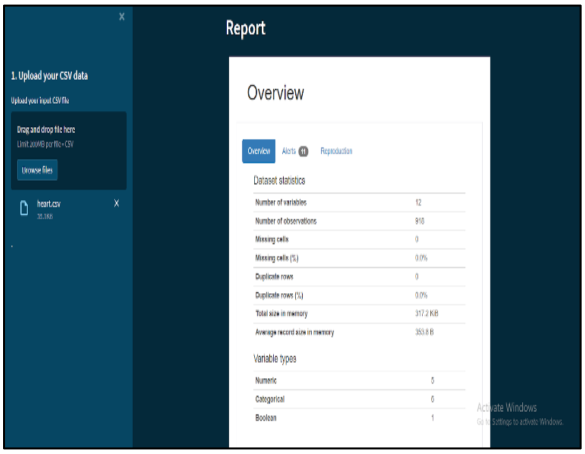
Cleaning of data is important before doing any modeling on the data to make sure that the data is credible and clean. Pandas is a well-known Python package that focuses on data analysis and manipulation.

4.1 Algorithm

Pandas DataProfiling uses EDA, also known as exploratory data analysis, to clean the data. Exploratory Data Analysis is frequently required when working with big amounts of data. We require a full explanation of the many columns available and their relationships, as well as information on null checks, data types, missing values, and so on. This is where the Python module of Pandas profiling performs the EDA and provides a full description with just a few lines of code. Exploratory data analysis (EDA) is the process of examining a dataset for patterns, similarities, and outliers (outliers), and then formulating hypotheses based on what we've discovered.

EDA comprises generating summary statistics for numerical data in the dataset and creating various graphical representations in order to better understand the data. It assists in the cleaning of data, such as missing numbers, and the usability of the data. The report shows information about cleaned data and the DataFrame. The application allows you to download the DataFrame as well as the report. By using EDA, the analysis of data and its parameters is done. This is the benefit of using EDA for Data cleaning.

Figure 1: Block Diagram for Data Cleaning



1. Upload your CSV data

Upload your input CSV file

Drag and drop file here
Limit 200MB per file • CSV

Browse files

heart.csv
35.1KB

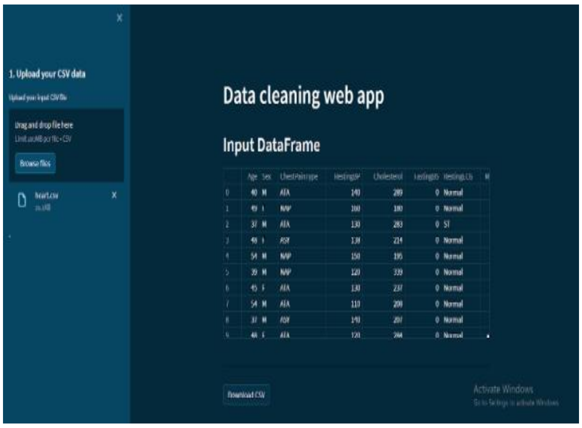
Dataset statistics

Number of variables	12
Number of observations	918
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	317.2 KiB
Average record size in memory	353.8 B

Variable types

Numeric	5
Categorical	6
Boolean	1

Figure 2: EDA for Data Cleaning



1. Upload your CSV data

Upload your input CSV file

Drag and drop file here
Limit 200MB per file • CSV

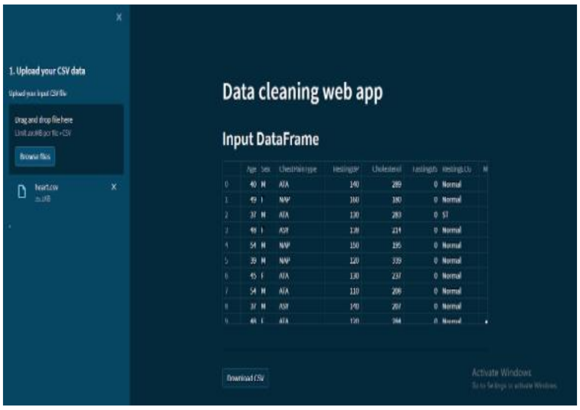
heart.csv
35.1KB

Data cleaning web app

Input DataFrame

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	M
0	40	M	ATA	140	289	0	Normal	
1	49	F	NAP	160	180	0	Normal	
2	37	M	ATA	130	283	0	ST	
3	48	F	ASY	138	214	0	Normal	
4	54	M	NAP	150	195	0	Normal	
5	39	M	NAP	120	339	0	Normal	
6	45	F	ATA	130	237	0	Normal	
7	54	M	ATA	110	208	0	Normal	
8	37	M	ASY	140	207	0	Normal	
9	48	F	ATA	120	284	0	Normal	

Figure 3: Cleaning of Data



1. Upload your CSV data

Upload your input CSV file

Drag and drop file here
Limit 200MB per file • CSV

heart.csv
35.1KB

Data cleaning web app

Input DataFrame

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	M
0	40	M	ATA	140	289	0	Normal	
1	49	F	NAP	160	180	0	Normal	
2	37	M	ATA	130	283	0	ST	
3	48	F	ASY	138	214	0	Normal	
4	54	M	NAP	150	195	0	Normal	
5	39	M	NAP	120	339	0	Normal	
6	45	F	ATA	130	237	0	Normal	
7	54	M	ATA	110	208	0	Normal	
8	37	M	ASY	140	207	0	Normal	
9	48	F	ATA	120	284	0	Normal	

4.2 Missing value handling

There are different ways in which missing values in the dataset can be handled.

- Identifying the missing value
- Replacing the missing value
- Dropping the missing value
- Filling the missing value,

However, in this application, the missing values are identified and replaced with zero. FOr identifying a missing value, there are different types of it and can be classified as

- Standard missing value
- Non-standard missing value
- Unexpected missing value

Standard missing value includes a blank cell or NaN. It can be detached by using the Pandas .isnull() method.

The Non-standard missing value will look like …, ___ or —, ?? na, etc. It can be identified by using the .unique() method on DataFrame.

Unexpected missing values generally have ?? X or wrong numbers like 99999. These are identified and the row is either dropped or replaced with NaN.

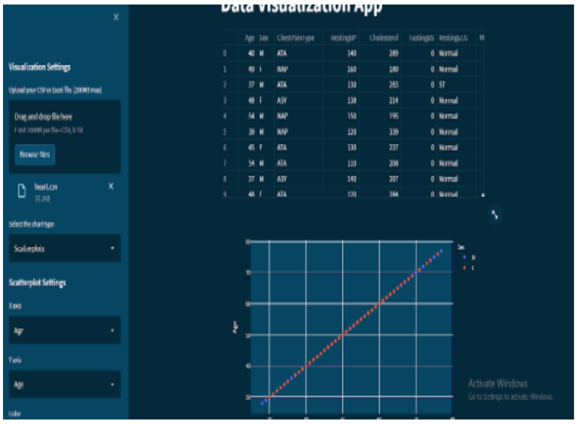
5.0 Data Visualization

It's a graphical depiction of facts or information. Cleaning the data using different cleaning algorithms in data science is required for proper display. Data visualisation tools make it simple to examine and grasp trends, outliers, and patterns in data by using visual features such as charts, graphs, and maps.

5.1 Line plot

Matplotlib is a Python data visualisation package. The pyplot library, which is a sub-library of matplotlib, is a set of functions that can be used to create a variety of graphs. Line charts are used to depict the relationship between two data points X and Yo on a distinct graph axis.

Figure 4: Line Plot



Data Visualization App

Visualization Settings

Upload your CSV or Excel file. (200MB max)

Drag and drop file here
Limit 200MB per file • CSV, XLSX

heart.csv
35.1KB

Select the chart type: Scatterplot

Scatterplot Settings

X axis: Age

Y axis: Age

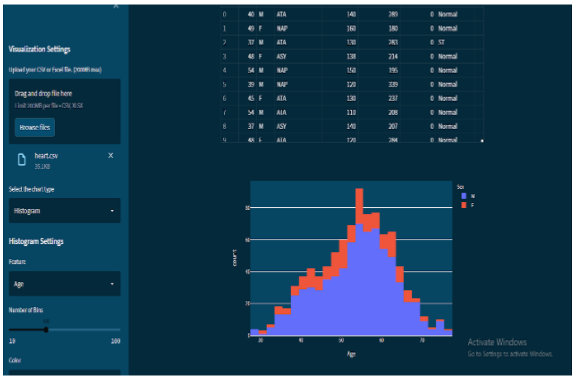
Color

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	M
0	40	M	ATA	140	289	0	Normal	
1	49	F	NAP	160	180	0	Normal	
2	37	M	ATA	130	283	0	ST	
3	48	F	ASY	138	214	0	Normal	
4	54	M	NAP	150	195	0	Normal	
5	39	M	NAP	120	339	0	Normal	
6	45	F	ATA	130	237	0	Normal	
7	54	M	ATA	110	208	0	Normal	
8	37	M	ASY	140	207	0	Normal	
9	48	F	ATA	120	284	0	Normal	

5.2 Histogram

A histogram is a graph that shows how data is divided into groups. It's a method of graphically showing numerical data distribution that's quite precise. In this type of bar graph, the X-axis shows bin ranges, while the Y-axis represents frequency. To make a histogram, divide the whole range of values into intervals and count the values that fall within each interval. Bins on a graph are defined as non-overlapping, sequential intervals of variables. Use the matplotlib.pyplot.hist() function to compute and build a histogram of x. The frequency is shown on the vertical axis of a histogram, and the horizontal axis represents another dimension. It usually has bins, each of which has a lowest and maximum value, as well as a frequency ranging from x to infinite.

5.3 Boxplot



Visualization Settings

Upload your CSV or Excel file. (200MB max)

Drag and drop file here

Limit 200MB per file • CSV, XLSX

heart.csv 35.1KB

Select the chart type: Histogram

Histogram Settings

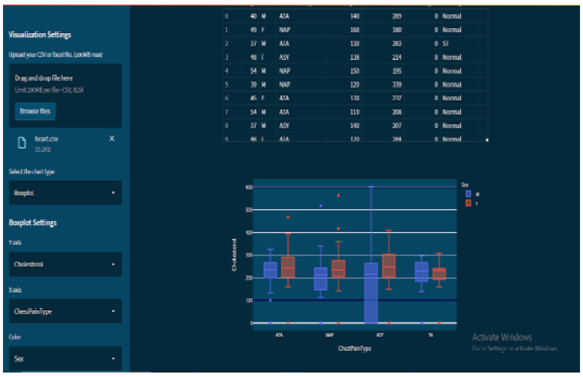
Feature: Age

Number of Bins: 40 (10–100)

Color

0	40	M	ATA	140	289	0	Normal
1	49	F	NAP	160	180	0	Normal
2	37	M	ATA	130	283	0	ST
3	48	F	ASY	138	214	0	Normal
4	54	M	NAP	150	195	0	Normal
5	39	M	NAP	120	339	0	Normal
6	45	F	ATA	130	237	0	Normal
7	54	M	ATA	110	208	0	Normal
8	37	M	ASY	140	207	0	Normal
9	48	F	ATA	120	284	0	Normal

Figure 5: Histogram



Visualization Settings

Upload your CSV or Excel file. (200MB max)

Drag and drop file here

Limit 200MB per file • CSV, XLSX

heart.csv 35.1KB

Select the chart type: Boxplot

Boxplot Settings

Y axis: Cholesterol

X axis: ChestPainType

Color: Sex

0	40	M	ATA	140	289	0	Normal
1	49	F	NAP	160	180	0	Normal
2	37	M	ATA	130	283	0	ST
3	48	F	ASY	138	214	0	Normal
4	54	M	NAP	150	195	0	Normal
5	39	M	NAP	120	339	0	Normal
6	45	F	ATA	130	237	0	Normal
7	54	M	ATA	110	208	0	Normal
8	37	M	ASY	140	207	0	Normal
9	48	F	ATA	120	284	0	Normal

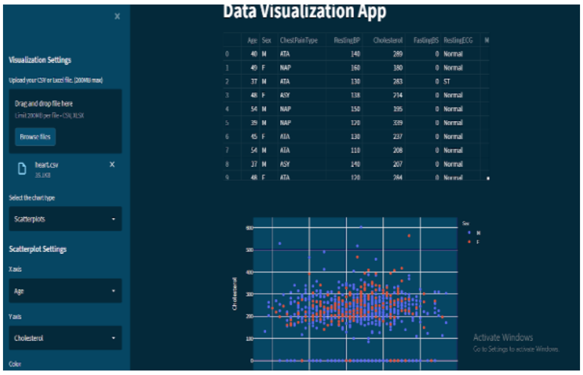
Figure 6: Box Plot

A Whisker Plot, also known as a Box Plot, is a graph that shows the minimum, first quartile, median, third quartile, and maximum values for a set of data values. In the box plot, a box is drawn from the first to the third quartile, with a vertical line crossing through it at the median.

The data to be plotted is represented by the x-axis, while the frequency distribution is represented by the y-axis.

5.4 Scatter plot

Scatter plots are used to visualize correlations between variables, with dots representing the link. To create a scatter plot, use the matplotlib library's scatter() method.



Data Visualization App

Visualization Settings

Upload your CSV or Excel file. (200MB max)

Drag and drop file here

Limit 200MB per file • CSV, XLSX

heart.csv 35.1KB

Select the chart type: Scatterplots

Scatterplot Settings

X axis: Age

Y axis: Cholesterol

Color

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	M
0	40	M	ATA	140	289	0	Normal	
1	49	F	NAP	160	180	0	Normal	
2	37	M	ATA	130	283	0	ST	
3	48	F	ASY	138	214	0	Normal	
4	54	M	NAP	150	195	0	Normal	
5	39	M	NAP	120	339	0	Normal	
6	45	F	ATA	130	237	0	Normal	
7	54	M	ATA	110	208	0	Normal	
8	37	M	ASY	140	207	0	Normal	
9	48	F	ATA	120	284	0	Normal	

Fig 7: Scatter plot

6.0 Data Prediction

Once the data is clean, it undergoes preprocessing and is used for prediction. The huge dataset is separated into two sections. They are known as the training sets and testing sets. The training set is the data which gets trained on 4 types of models. Then using the test dataset, one can check if the trained model is working properly and the predict outcome is accurate or not. There are 4 types of models we will be building: they are based on Decision Tree, Naive Bayes, SVM and Random Forest algorithms.

6.1 Decision tree

Decision tree algorithm is one of the earliest and oldest machine learning algorithms of all time. A decision tree model checks out for the logic and respectively provides an outcome for classifying data elements into a tree-like structure. This algorithm can be applied to the specified set of inputs and outputs and can also solve classification and regression problems. It consists of multiple levels of a group of nodes branched and each node is considered data through which to make a decision. Most of the time the first knot is the best feature and is called root while the last knot is called paper. This algorithm is also easy and simple yet very effective.

6.2 Naive bayes

Naïve Bayes is a classification algorithm that works by using the Bayes' theorem. This logic can describe the probability of an event solely depending on the previous knowledge of conditions related to that very same event. An assumption is made that a specific attribute or feature in a dataset is not directly related to any other attributes. Hence these attributes also known as features for that particular dataset could have interdependence among themselves. This classification can solve problems related to diagnosis as well as forecasting because it uses a basic model that allows decision-makers to capture the uncertainty in the experiment on the dataset

6.3 SVM

The SVM or the Support Vector Machine algorithm can be used for the classification and prediction of linear as well as non-linear data. Initially, it maps each dataset point into an n-dimensional space. Here, n is the number of features or attributes of that particular given dataset.

Then it creates the hyperplane that separates the data points into two different categories while maximizing the marginal distance for both the categories. Also it tries to minimize the classification errors. For prediction, we need to get the hyperplane that differentiates between the two categories by using the maximum margin.

6.4 Random forest

A random forest algorithm is a classifier that makes use of more than one Decision Trees.. In the Random Forest Algorithm, each of the trees present a different part of that input given. Each of them gives an outcome. The algorithm then chooses the prediction which either has the most output or calculates the average of all the outcomes of the decision trees in the forest.

7.0 Health Monitoring

Health tracking may be very crucial in phrases of prevention, specifically if the early detection of sicknesses can lessen struggling and scientific costs. The analysis and activate remedy of diverse sicknesses can significantly enhance options for the scientific remedy of the affected person.

So the Internet of Things (IoT) primarily based totally fitness tracking machine is the modern answer for it. Remote Patient Monitoring association empowers remark of sufferers outdoor of standard medical settings (e.g. at home), which expands get admission to to human offerings workplaces at bringing down expenses. The middle goal of this venture is the layout and implementation of a clever affected person fitness monitoring machine that makes use of Sensors to song affected person fitness and makes use of the net to tell their cherished ones in case of any troubles SMS primarily based totally affected person flourishing viewing and IOT primarily based totally affected person checking framework. The middle goal of this venture is the layout and implementation of a clever affected person fitness monitoring machine. Our machine constantly video display units affected person's critical symptoms and symptoms and senses abnormalities. Thus, reduces the want for guide tracking performed via way of means of the scientific staff.

7.1 Hardware description

NODEMCU ESP8266 Board: Node MicroController Unit, Serial WIFI Wireless Transceiver Module is a self-contained SOC with incorporated TCP/IP protocol stack which can provide any microcontroller get entry to in your WiFi network.

The ESP8266 is able to both web website hosting an software or offloading all Wi-Fi networking features from any other software processor.

AD8232 ECG Sensor: An electrocardiogram (ECG) is a easy check that may be used to test your coronary heart's rhythm and electric interest. Sensors connected to the pores and skin are used to hit upon the electric indicators produced with the aid of using your coronary heart whenever it beats.ECG Development Kit also can be applied to monitor (non-invasive) floor EMG, presenting a illustration of the muscle interest on the size site. Combined with Shimmer's included 9DoF inertial + altimeter sensor platform, extra context may be given to the wearer's interest and situation in actual time.

MAX30100 Pulse Sensor: Pulse oximetry is a noninvasive take a look at that measures the oxygen saturation degree of your blood. It can unexpectedly discover even small modifications in oxygen levels. These levels display how efficiently blood is sporting oxygen to the extremities furthest out of your heart, which includes your fingers and legs.

7.2 Patient monitoring process

Pulse oximetry is a noninvasive test that measures the oxygen saturation diploma of your blood. It can all of sudden find out even small changes in oxygen degrees. These degrees show how correctly blood is carrying oxygen to the extremities furthest from your coronary heart, which incorporates your arms and legs. A programmed faraway fitness watching device is applied to quantify a affected person's frame temperature, pulse with the aid of using making use of implanted innovation. These sensors in general encompass watching the fitness condition, fall detection and sleep sample of the affected person. Heart beat sensor, Body temperature sensor, and blood oxygen level (MAX30100). These sensors paintings autonomously with every other. The measured studying from the sensor is damaged down for the affected person and is made available to the professional or to any worried character withinside the kind of the net or clever phones.

This net interface's flexible software serves because the person interface for this model. The sensor modules fused withinside the implanted tool yields the automated cost which may be interpreted with the aid of using aligning the sensors. These readings are transmitted to the Firebase and displayed at the website.

7.3 User interface

React App Interface*:* The Web software gives the person with an interface to engage with the tool. The software gives the person actual time studying of the sensors thereby getting the affected person's popularity to the clients.

The software queries facts into the database and shows it. Heartbeat, SPO2 are shown. The React net software indicates the measured parameters coronary heart rate, SPO2.

FireBase*:*The Firebase Realtime Database is a cloud-hosted NoSQL database that helps you to keep and sync facts among your customers in actual time. FireBase gives authentication and cease- to-cease encryption at some stage in all factors of connection, in order that facts is by no means exchanged among gadgets and FireBase with out confirmed identity. Thus, facts is securely being transmitted to the Firebase platform.

Figure a: Patient Monitoring Modular Diagram

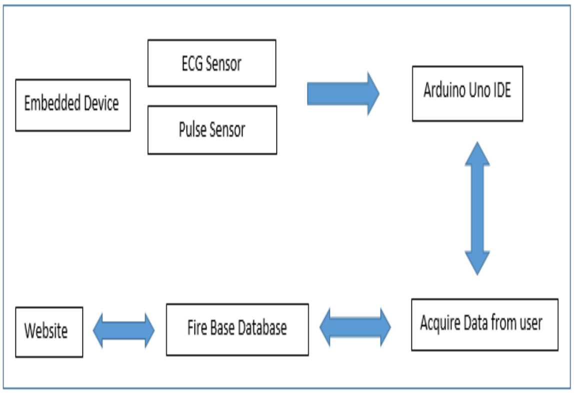
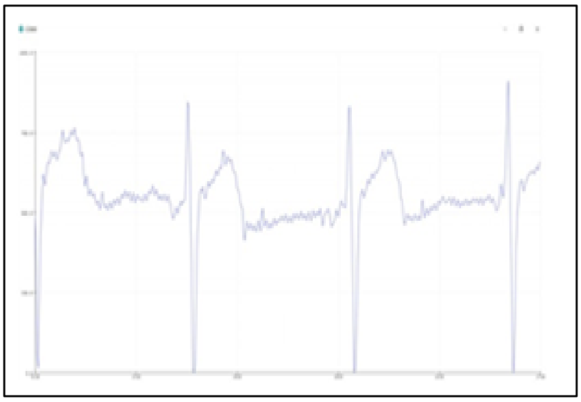


Fig b: ECG Graph



8.0 Discussion of Result

Before you jump into the result we got and the accuracy we obtained from the different prediction algorithm. Let us discuss the dataset used and the different attributes of the same.

8.1 Properties of the dataset

In the dataset we have used there are 11 attributes according to which we can predict if a particular person may or may not have a heart disease. Here is a detailed table of the different attributes or features of the dataset and the description of them.

8.2 Comparison between different models

As discussed previously, four algorithms were used for prediction, that is Decision Tree, Naive Bayes, SVM and Random Forest. Here is the accuracy of each with and without our data cleaning model:

Table 1: Dataset Attributes

Features	Description
Age	Numeric Value
Sex	Male or Female
Chest Pain	AST, ATA, NAP, TA
Resting Blood Pressure	Numeric Value
Cholesterol	Numeric Value
Fasting Blood Sugar	0 or 1
Resting ECG	LVH, Normal, ST
Max Heart Rate	Numeric Value
Exercise Angina	Yes or no
Old Peak	Numeric Value
ST Slope	Down, Flat, Up

Table 2: Comparison of trained models

Algorithm	Without Cleaning	With Cleaning
Decision Tree	0.784	0.828
Naive Bayes	0.875	0.870
SVM	0.859	0.882
Random Forest	0.885	0.899

9.0 Conclusion

Cleaning, prediction and visualization of the given disease data, one can check if they will be affected by the disease or not and our user-friendly report. The early prediction of disease before its onset in a patient gives clinicians additional lead time to plan and execute treatment plans.

It is observed that our protocol can be used for various applications related to healthcare and patient monitoring such as heart disease prediction or cancer severity classification. Patient health monitoring system based on IoT uses the internet to effectively monitor patient health and helps the user monitor their loved ones from work and saves lives.

10.0 Acknowledgement

We would like to express our gratitude to our professor Amit Hatekar and our parents for their immense support and guidance.

References

[1] Joseph M. Hellerstein, “Data Cleaning,” EECS Computer Science Division UC Berkeley, 2008.

[2] Dataset, Heart-UCI-EDA Kaggle.

[3] Johnathan Wanderer, “Clinical Data Visualization: The Current State and Future Needs” 2016.

[4] Hyejung Chang “Interactive Visualization of Healthcare Data Using Tableau”, Healthcare Informatics Research.

[5] Charlotte Prins, Gijs Van Pottelbergh, Pavlos Mamouris, Bert Vaes & Bart De Moor “An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge” 2020.

[6] Robert Hoyt (Author), Robert Muenchen (Author) “Data Preparation and Exploration: Applied to Healthcare Data” 2018.

[7] Hu, Yuanzhang MDa; Yu, Zeyun, Cheng, Xiaoen MDa,∗; Luo, Yue, Wen, Chuanbiao MDa “A bibliometric analysis and visualization of medical data mining research” 2020.

[8] Ying Yang, Xinxiang, “Analysis and Visualization Implementation of Medical Big Data” 2020.

[9] Nada Lavrac, Marko Bohanec, Aleksander Pur, Bojan Cestnik, Marko Debeljak, Andrej Kobler, “Data mining and visualization for decision support and modeling of public health-care resources” 2018.

[10] Alan Pryor, Ph.D., Kaggle publish Data Notebook “Detailed Cleaning/Visualization (Python)” 2015.