Cardio-Vascular Disease prediction using Machine Learning Algorithms

Terrance Thomas
National College of Ireland
School of Computing
Dublin, Republic of Ireland
x18184928@student.ncirl.ie

Abstract—**Cardio-Vascular diseases (CVDs) ranks first for global death. 17.9 million deaths are due to CVDs each year as per world health organization. CVDs are a group of diseases or disorder of blood vessels and heart. CVDs include coronary heart disease (CHD), rheumatic heart disease (RHD), cerebrovascular disease and others. Every 4 deaths of 5 CVDs are due to heart attacks and strokes. Also, one-third of those deaths occur to people below the age of 70 years. In the modern age, with such large medical and healthcare data available machine learning (ML) can be used effectively for predicting and making informed decisions. This prediction can be used to provide tailored treatments and care.**

*Keywords—**Cardiovascular disease, Prediction, Data mining, Heart Disease, Heart Attack, Machine Learning.***

I. Introduction

Heart disease or cardiovascular syndrome is a type of disease that happens due to difficulty in the blood vessels, veins, and arteries which are connected to the heart and other body parts [1].

A few of the most common factors that causes CVD are no physical activity, alcohol consumption, no proper diet, use of tobacco, high level of bad cholesterol, obesity, high level of blood sugar, etc. These factors are called as controllable factors as these can be changed over time from a person and not acquired. Also, other factors lead to CVD such as age, sex, family history. These factors are uncontrollable factors. The controllable and uncontrollable factors are the 2 major risk factors for CVD [1].

The growth of blood clots is increased by the chemical substance present in tobacco. If a person is obese then the chances of heart problems also increase especially to people who have extra fat on the waist. Unwanted cholesterol present in the body builds up a wall in the arteries leading to atherosclerosis, which is again a type of heart disease. Diabetes a life-threating and lifelong condition increase the blood pressure and cholesterol level as well [1].

Medical science is regarded as the most emerging field for research. This field is very critical to deal with because the information collected may not be correct or other complicated factors that need to be diagnosed in a different way [2]. Data mining and ML are used by numerous researchers to support the health care industry. It plays a key role in early forecasting predicting systems [1]. There are multiple algorithms and methods that can be used for the predicting system. Hidden patterns and relationships can be extracted using data mining from large data sources. This helps the doctors to make a very informed and efficient decision [3].

About 25% of death between the age group of 25 to 69 years are because to heart diseases. If this trend continues, it is estimated that by 2030 sixty percentage of death will be due to CVD [1]. Hence it is very important to keep updating the precision of the detection using new and combination of methods.

The data used for this research project is from Kaggle. The dataset consists of 70,000 rows and 14 columns. The columns are id, age day, age year, gender, height, weight, High blood pressure, low blood pressure, cholesterol, glucose, smoking, alcohol consumption, active and cardio. Gender has values 1 and 2 where 1 is for men and 2 is for women. Cholesterol has 3 levels where 1 is normal, 2 is above normal and 3 is high. Smoking and alcohol both are binary values where 0 is no and 1 is yes. Active is also consists of binary value where 0 is an inactive lifestyle and 1 is an active lifestyle. The final variable is the result cardio which is again a binary value where 0 is CVD is no and 1 is CVD is yes. The data is nonbiased data and overcomes different types of bias. The selection of test and train will be done randomly by the split feature. The information provided as the result of CVD almost has an equal number of cases.

II. Project Goal

The primary goal of the project to predict if a person is likely to have CVD or not. However, the most important factor in the prediction is to have high accuracy and to reduce the cases of false positive, as in this case a person is predicted to not have CVD while he/she has CVD.

III. Ethical Concerns

The general data protection regulation (GDPR) is a European regulation on data protection and privacy. The most significant impact is on internet users. GDPR was implemented to connect the legal gap with emerging technologies. Data protection is related to data share and access [4].

The data used for research is taken from Kaggle which is a subsidiary of Google and a community for data scientist and ML practitioners. Additionally, the dataset used has no information that can reveal the identity of the user. It consists of some basic information and some behavioural data of the user. Each patient has been assigned an id to keep the user anonymous. No biometric data of users have been used for the project.

The data used aims to predict CVD with higher accuracy and indeed help the medical research field for the wellbeing of humans and not for any unethical purpose.

IV. Strategy

The project aims to predict CVDs with better accuracy. Additionally, early detection is as important as well. This directly results in increasing the life span of people who are likely to have CVDs.

A. *Business strategy*

In the modern age, there are many options available for healthcare. Also, the ever-developing technology makes it more competitive as the competitors try to be better to compete and survive in the market.

1) Patient satisfaction: This refers to the entire experience that the patient has since they arrived at the healthcare facility. Right from the start, the patient should feel as they are given top priority. For this, the staff has to be trained on how to handle a patient and how to engage with them. Such a small factor considerably changes the attitude of the patient. Also, before any decision made the patient should be well informed about it. The journey and recovery process is a crucial part. Finally, when the patient should be given a feedback form. This will help to understand where and which field the work has to be done.

2) Employee Engagement: This includes listening to employee issues and figuring out a win-win situation. If the employee issues are not addressed they can always move to the competitors. Hence, it is important to always have timely feedback from the employees and try to solve their grievances. Additionally, this includes having employees enrolled in leadership programs, soft skill programs where they can grow as an individual and over the time climb the corporate ladder as well.

3) Patient Education: This is another key factor. Although everything is available on the internet today, yet it is very important to educate the patients on how they can prevent such a situation in the future. Also, the patients should be educated about their condition.

4) Patient user experience: This is in regards to the online experience. The website and mobile application should be developed in such a way that the patients or the person who is trying to search for information is provided accurately and with ease. This can be information related to a disease, condition or even general information. Additionally, chatbots can be used to make the user journey easy.

5) Collaborate with health plan providers: There are multiple health plans available and a user can select the health plan which is affordable and covers the conditions they want to. Hence, it is very important to have a collaboration with not only major but with majority health plan providers to have a seamless journey for the patient.

B. *Marketing strategy*

Marketing is a key function in any organization. Marketing strategy is important to make the world know about the organization, to compete with competitors and to get more people. A different marketing strategy can be used depending on the phase of the organization.

When the organization is new, the main aim is to make the people aware of such an organization o brand. This strategy does not have the best conversions but gets you a lot of audiences. Online and offline modes can be used to make a brand famous.

Now once the name is out there, the next step is to get the return on investments. Start schemes like referral schemes, discount schemes, etc. This will get a lot of people to healthcare. Patients now having a seamless experience is very important. Only then they can be retained.

Converting a new customer takes far more monetary investment than converting an existing customer. So, every patient must have the best experience. Also, if they have a seamless experience, they would spread good reviews about the place to relatives and friends and this is the best marketing with almost zero investment.

C. *Insights*

This is another important factor in the strategy of the organization. This is where the organization can use all the data gathered overtime to make the place better in places where it is lagging. This can either be from the patient's feedback about the service and experience or can be the outcome of a marketing strategy. Based on the result a calculated decision must be taken on how to improve the service and user experience. Also, which marketing strategy to be used and how much to invest in the new scheme must be decided.

V. VISUALIZATION

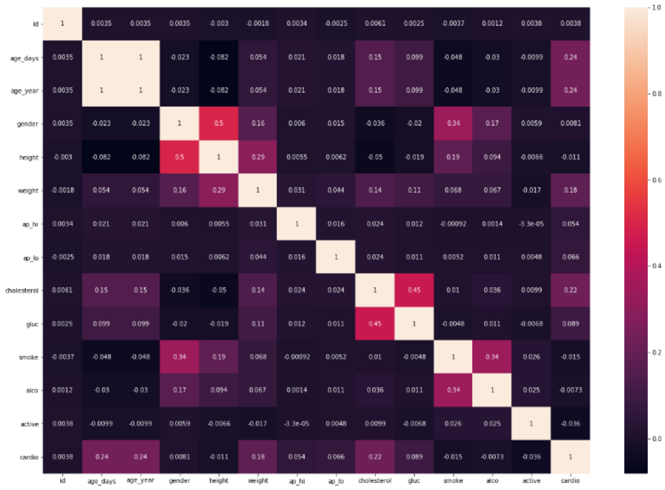

Figure 1: Correlation Matrix

Figure 1 shows the coefficient correlation between the variables. Higher the score the more they are correlated the variables are. This matrix helps in determining which variables to be used while applying the ML method to get better results. The variables can be positively or negatively correlated with (1 or -1 being the highest value and zero the least value).

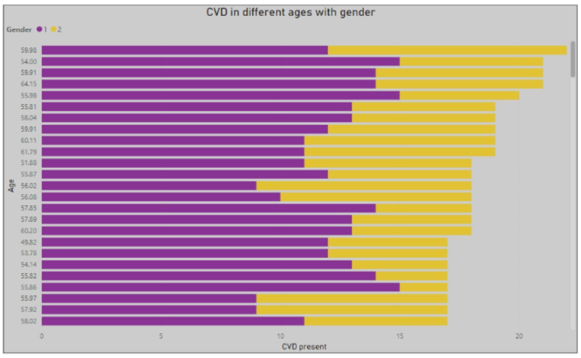

Figure 2: Stacked Bar Graph

Figure 2 shows the bar graph and how much patients have CVDs with their gender across different ages. The graph is sorted in descending order. So, the number of patients in an age with CVDs are kept on the top. In this case, the people with the age of 59.98 years have the highest occurrence of CVD. The purple color is patients who are male while yellow color is for female patients >.

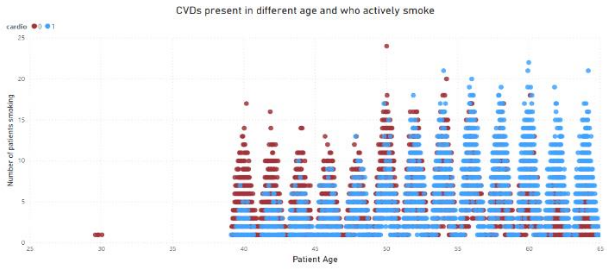

Figure 3: Scatter Chart

Figure 3 shows the scatter chart and how many patients have CVDs and how many don't have CVDs with their gender across different ages who actively smoke. The graph is sorted in descending order. The red color is patients who don't have CVDs while blue color is for patients who have CVDs.

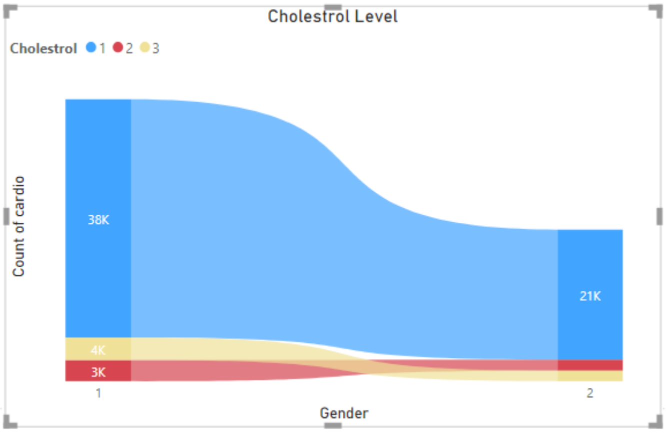

Figure 4: Ribbon Chart

Figure 4 shows the ribbon chart and different levels of cholesterol among men and women. The blue color is for level 1 which is a normal level. Red is for level 2 which is above the normal level. Yellow is for level 3 which indicates a high cholesterol level.

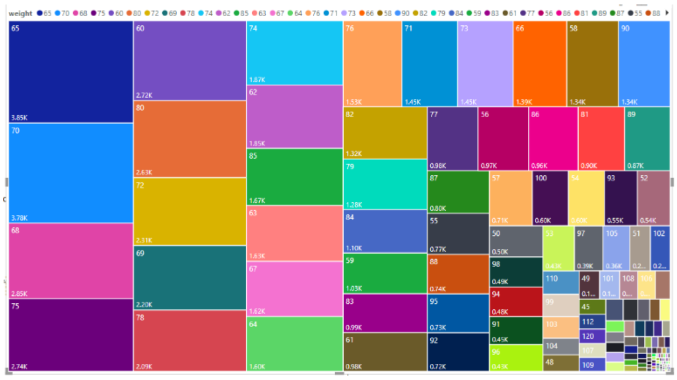

Figure 5: Tree Map

Figure 5 shows patients of different ages. The figure explains the number of people who are physically active and have CVDs.

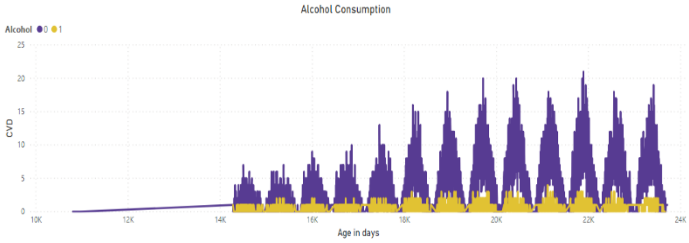

Figure 6: Line Chart

Figure 6 shows alcohol consumption among the patients by their age in days. Purple color is patients who do not consume alcohol while yellow is for patients who consume alcohol.

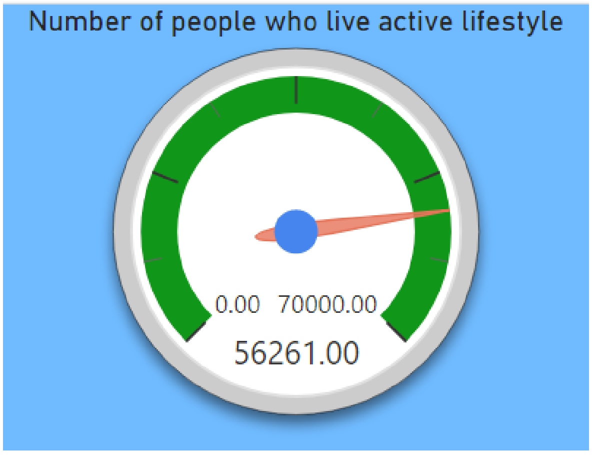

Figure 7: Dial Gauge

Figure 7 shows the number of patients from the dataset who have an active lifestyle.

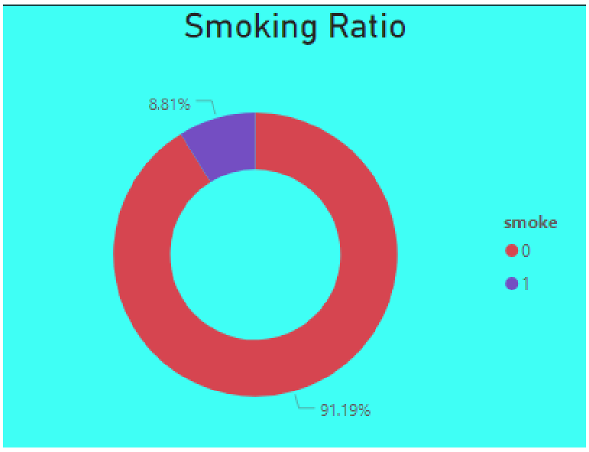

Figure 8: Donut Chart

Figure 8 is a donut chart representing the smoking ratio in the dataset. Red color indicates non-smokers while purple represents smokers.

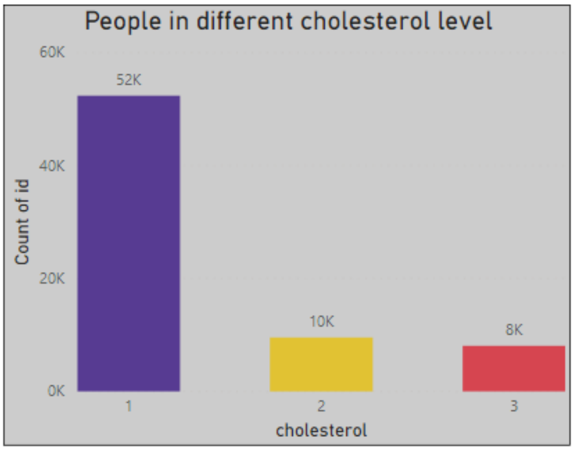

Figure 9: Column Chart

Figure 9 shows the 3 categories of cholesterol levels and the number of patients at each level. Purple is level 1 which is normal; followed by yellow which is above normal level while red is high level.

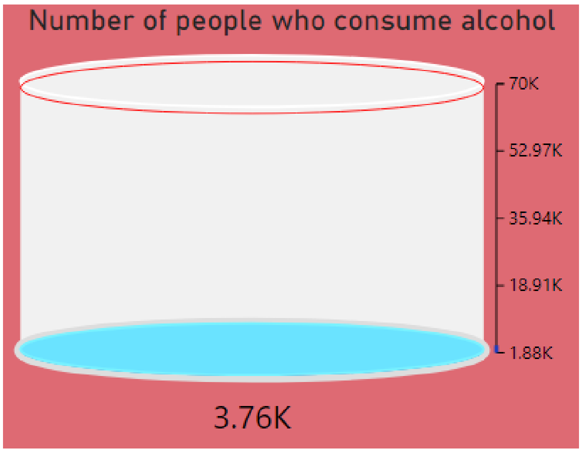

Figure 10: Cylindrical Gauge

Figure 10 shows the number of patients who consume alcohol from the dataset. About 3.76k people consume alcohol of a total of 70k.

VI. APPICABLE TECHNIQUES

As CVDs is one of the leading cause of deaths over the years, a lot of research has been done over the years. Various techniques, datasets, various methods have been applied tried and tested. However, unless the data collected is not correct the prediction made cannot be correct. So, an error in data collection or the devices can result in the wrong prediction and if the false positive percentage is high it will lead to CVDs not being detected and the doctors may not treat the patient in early stages and ultimately increase the death toll.

K-means: K-means clustering is a strategy for vector quantization, initially from signal processing, that is mainstream for group investigation in information mining. K-means clustering intends to partition n observations into k groups in which every observation has a place with the bunch with the closest mean, filling in as a model of the group.

Decision Tree (DT): DT is the most remarkable and mainstream tool classification and prediction. DT is a flowchart like a tree structure, where each inside node indicates a test on a property, each branch represents a result of the test, and each leaf node (terminal node) holds a class label.

C4.5: C4.5 is a calculation used to create a DT. It tends to be utilized for order, and C4.5 is regularly alluded to as a measurable classifier.

Neural network (NN): NN is an information processing structure, parallelly disseminated, comprising of various quantities of handling components in a particular node. Unidirectional signal channels called connections to interconnect them. Each handling component has a solitary yield association stretching into numerous associations. Every conveys a similar sign, for example, the handling component yield signal. The NN can be characterized in two fundamental gatherings as indicated by the way they learn.

Naive Bayesian: It is a probabilistic ML model that’s used for the classification task. It is based on Bayes theorem. The formula for Bayes theorem is as follows:

P(A/B) = (P(B/A) x P(A)) / P(B).

Random Forest (RF): It comprises of different DT. It utilizes bagging and feature randomness when constructing every individual tree to attempt to make an uncorrelated forest of trees whose prediction is better to any individual tree

Principal Component Analysis (PCA): It is a statistical procedure that uses orthogonal transformation, It converts a set of observations into a set of linearly uncorrelated variables. This is called the principal component. The first principal component has the highest variance after the transformation.

The authors use the dataset of patients that show various factors contributing to CVDs. Two ML algorithms were applied on the dataset viz K-means clustering classification and DT. Multiple measuring factors like precision, recall, accuracy and F measure were used to decide which algorithm had a better model. Mean absolute error, Kappa statistics and Root mean square error were also considered. The authors suggested using other factors that were not considered for this research, as these factors can be a vital factor for CVD prediction [1].

The research in this paper a framework is developed for the clinical diagnosis. C4.5, NN, and naïve Bayesian algorithms are used for the prediction of heart disease. The C4.5 helps to build the DT which is used in the research as well. In this research even though the data set was large, the results obtained were not significant. It can be due to the incompleteness of the data; however, the research shows how multiple algorithms can be combined to obtain results [5].

The authors use a fuzzy logic system to predict CVDs. Heart rate is considered in this dataset where they are classified into 4 groups. Other details like chest pain, blood pressure, cholesterol, blood sugar, gender, and age are considered for the research. Even tough fuzzy logic is the most popular method to use, the system created was by experts and reduces time in prediction [6].

The research uses multiple ML algorithms like Naive Bayesian and RF for prediction. The aim of the research is to predict CVD in people under the age of 50. PCA has been used in the research as well [7].

VII. References

[1] S. S. a. S. D, "Early Forecasting and Prevention of Cardio Vascular Disease Based on Human Life Style Factors," in *Conference on Emerging Devices and Smart Systems (ICEDSS)*, Tiruchengode, 2018.

[2] V. M. a. A. Goyal, "X-Cardio: Fuzzy Inference System to Diagnose Heart Diseases," in *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida (UP), India, 2018.

[3] A. S. a. S. S. C. Suvarna, "International Conference on Computing Methodologies and Communication (ICCMC)," in *Efficient heart disease prediction system using optimization technique*, Erode, 2017.

[4] S. O. a. M. I. M. Karampela, "Exploring users' willingness to share their health and personal data under the prism of the new GDPR: implications in healthcare," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019.

[5] I. P. a. S. S. K. Rajmohan, "Prediction and Diagnosis of Cardio Vascular Disease -- A Critical Survey," in *World Congress on Computing and Communication Technologies*, Trichirappalli, 2014 .

[6] L. A. a. I. B. A. Duisenbayeva, "Using Fuzzy logic concepts in creating the decision making expert system for cardio — vascular diseases (CVD)," in *IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, Baku, 2016.

[7] D. K. a. B. Priyadharshini, "Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using classification techniques within fifty years of age," in *2nd International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, 2018.

.