Analysis of Healthcare Expenses DAP Group Project

Noel Joseph
Data Analytics
National College of Ireland
Dublin, Ireland
x18184499@student.ncirl.ie

Kamesh Munuswamy
Data Analytics
National College of Ireland
Dublin, Ireland
x19199562@student.ncirl.ie

Siddharth Shukla
Data Analytics
National College of Ireland
Dublin, Ireland
x19197403@student.ncirl.ie

***Abstract*—Health care sector of the U.S.A is predominantly owned by the private sector businesses. The annual expenditure of USA on healthcare per capita is $9,403 and if you convert this amount into GDP percentages, then in 2014 census, US spent 17.1% of the total GDP on Health Care which is equivalent to $3.65 trillion, an amount bigger than the total GDP of almost 50% of all the countries in the world. When it come to the medical facilities and services, it becomes very complex, as it is a big mesh of public and private funded patch work of individual systems and programs. The Americans are covered by both public and private enterprises with the majority of the part taken by the privately owned employers. The big chuck of this money comes from the Heart conditions. Not only the heart condition effects the health of an individual, it takes a heavy toll on his finances as well. Just the cost of cardiovascular disease treatment in the US in2014 was $444 billion. Similarly, there is a big disparity in the medical bills of people who smoke and non-smokers, since smoking is one of the reasons for a heart attack therefore indirectly contribution to more medical expenses.**

Keywords-Visualization, Forecasting, Statistical Model, Swarm plot, Medical Expenses.

I. INTRODUCTION

It is necessary to diagnose a disease at an early stage. This group project's main purpose is to develop a good understanding of the datasets which are generally taken in medical field. The understanding of data alongside the different characteristics of disease is important for developing/modelling a proper code for measuring the accuracy from the selected datasets. The papers that have been consulted for this analysis talk about the use of clustering and the simple statistics which apply over the heart related data sets to answer the questions related to the medical conditions for a comparative analysis. We will be harnessing the powers of data visualization and predictive modeling to analyze the healthcare costs a likeability of having a heart condition for given time duration.

The objective is to analyze each dataset individually and then try to find similar pattern or trend that can connect the dots between each other dataset that may result in a cumulative analytical conclusion. Different analysis and predictive modelling will also be used on each market which will be addressed individually and in detail in the methodology section. Then, we separately analyze and find the trend and similar pattern between all three-stock exchanges for a given time period.

II. RESEARCH QUESTION

1.What are the medical expenses distribution among Smokers and Non-Smokers?
2.What is the average cost distribution of different

drug/medical procedures in different hospitals across all the states in USA?
3.How do the different health aspects change the likeability of having a heart disease?

III. RELATED WORK

Following research work have been studied to form the base of our project:

[1] StevenB. Cohen et al research offers quantitative estimates of healthcare spending concentration and duration rates in the United States. Analyzes are used to find out the most important factor that helps to forecast the probability of experiencing high rates of medical expenditure in the following year.
[2] Ker-Tah Hsu et al paper aims to use the Grey Method for forecasting different forms of medical expenses. And then we seek to determine from that which one of them plays the decisive role in the financial distress of the National Health Insurance of Taiwan.
[3] IN BabakSohrabi et al paper Useful patterns from these large data sets are discovered and extracted to identify secret and worthy patterns for decision taking. This paper, too, aims to demonstrate the potential of the data-mining method to enhance the quality of decision-making in the pharmaceutical sector
[4] In Bibi Amina Begum et al research the author is Predicting thyroid disease using various classification methods. The main goal is to identify diagnosis of diseases with greater precision in the early stages.
[5] Sushmita Roy Tith et al intend to develop a model using supervised machine learning, and by analyzing it can find anomalies in the ECG report. To differentiate between regular and abnormal ECG, they used six supervised machine learning algorithms.
[6] Edvinas Narbutas et al paper talks about Creation of software environment tools. Presenting visualized graphs of the generated software data structures and examples. They have also introduced running applications on an Android mobile system and also suggested Opportunities for new modifications of computerized cardiovascular analysis systems to track long-term heart rate processes in the mobile world.
[7] The Judith M. Katzenellenbogen1 et al author examined the provision of guideline-recommended services for the management of acute rheumatic fever (ARF) and rheumatic heart disease (RHD.
[8] Mustafa Jan et al study tries to screen the medical databases and predictive modeling using soft computing techniques is considered a valuable and cost-effective choice for medical practitioners. They propose an Ensemble model method for combining the predictive capacity of a set with multiple classifiers to boost accuracy.
[9] S. Mohan et al proposed a novel approach designed to identify significant features by applying machine-learning techniques that improve cardiovascular disease prediction accuracy. The predictive model is implemented with different combinations of features and functions A range of proven classification techniques.
[10] In Davide Golinelli et.al the main objective of this study was to examine whether and how the report contained public health expenditure per capita (PHE) Results which are important metrics to be used in resource allocation decision-making processes.

IV. METHODOLOGY

This section of the report will tell us about the overall flow of the project. We have specially made the flowchart which gives a generalized view on the different steps that were undertaken in this project.

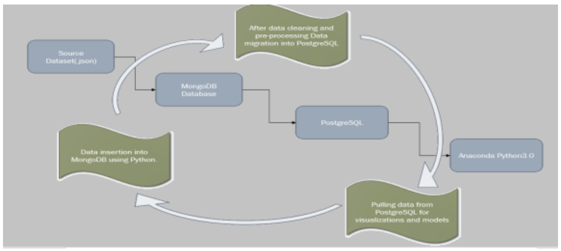

Source Dataset(.json)
MongoDB Database
PostgreSQL
Anaconda Python3.0
After data cleaning and pre-processing Data migration into PostgreSQL
Pulling data from PostgreSQL for visualizations and models
Data insertion into MongoDB using Python.

Fig 1. Project Development Methodology flow diagram.

- **Dataset Selection:** After going through the different dataset repository, we selected the three datasets.
- **Storing File in MongoDb:** The Json file is then stored into the MongoDb database using the python connection.
- **Data Cleaning:** We then go ahead and clean the data using Python. Eg: The id column which is auto- generated and that needs to be removed.
- **Target Table creation:** PostgreSQL connection is set up and the cleaned data is stored in the target SQL Database.
- **Modelling and Visualization:** The final Target table is pulled from PostgreSQL using python and finally modelling and Visualizations well be made.

Dataset insights:
Find below the dataset information that have been chosen for the project.

Dataset 1 - Medical Cost personal dataset

This Medical cost personal dataset was extracted from Kaggle.com. This dataset contains the information about the health measuring factors and the individual medical insurance cost. The dataset has total of 1339 records with 7 attributes such as age, sex, bmi, children, smoker, region, charges.

Link: https://www.kaggle.com/mirichoi0218/insurance

Dataset 2- Heart disease dataset

This dataset is related to heart disease which is also extracted from Kaggle.com. This heart disease dataset holds a complete set of information regarding the heart disease measuring parameters along with a "target" field referring to the presence of heart disease in patient. It is encoded as 0 = no heart disease and 1 = heart disease. There are total of 1026 records in this dataset along with 14 attributes. Link: https://www.kaggle.com/johnsmith88/heart-disease-dataset

Dataset-3 Consolidated Medical procedures/Drugs cost across US

The dataset has a custom-made repository of medical records that have been handpicked from major medicinal facilities from all the states of USA. It contains the various procedures and drugs that are available for a patient. The file holds more than 85000 rows and 12 columns.

Link:https://www.kaggle.com/speedoheck/inpatient-hospital-charges

Data pre-processing and EDA

All chosen three datasets related to health care are pre-processed individually and in-depth EDA has been performed. Based on the EDA, data were sharpened and reshaped for clear data visualization. Dataset-1 and Dataset-2 were merged into one singleton dataset using SQL join query when retrieving from PostgreSQL database. After which data visualization is carried out using python. Dataset-3 was pre-processed individually and data visualization is carried out using SQL queries to retrieve the data from PostgreSQL.

V. DATA VISUALIZATIONS

Medical expenses – Dataset 1

Data visualization based on Smokers

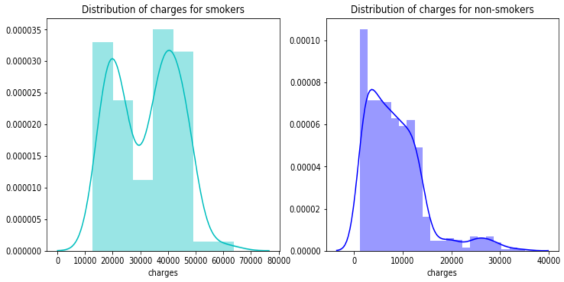


Fig 2: Smoker Vs non- Smokers based on medical charges.

The above chart is plotted against the smokers, non-smokers and medical charges. This is a distribution plot of smokers and non-smokers which gives the overall idea about the medical expenditure of smokers and non-smokers. From the left-hand side graph of smokers, it is been observed that smokers spend more than 60000 dollars for a time period. Wherein, Non-smokers spends around 30000 dollars on their health. From this graph we can infer that smokers have a tendency to garner higher medical expenses as compared to the non- smoking population.

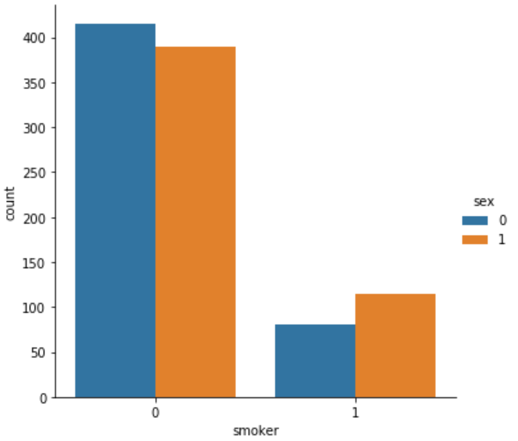


Fig 3: Smoker vs non-smoker based on gender

This bar graph is plotted against the smokers, non-smoker against the gender. Here, in the plot data is encoded as smokers - 1 and non-smoker - 0 while sex (male- 1 and female- 0). This bar graph is plotted to check the number of smokers and non-smokers among male and female. As we can see that, there are good number of non-smokers in both male and female. Perhaps, unsurprisingly male gender is prone to smoking habit having a high count than females.

Medical expenses vs Gender (male and female) smokers and non-smokers

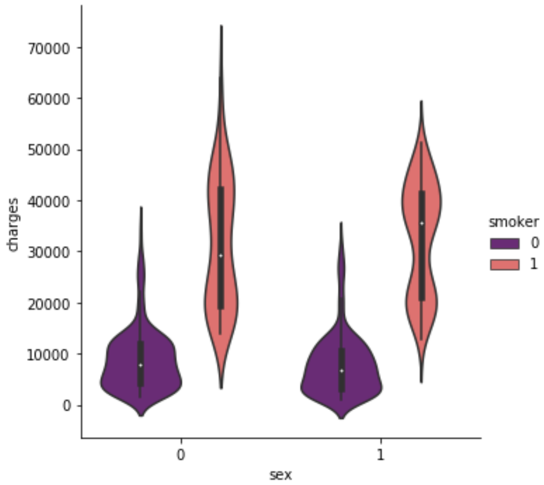


Fig 4: Medical expenses vs Gender

The above violin plot is plotted between gender (male-1, female- 0) smokers and medical expenditure. From this plot we can clearly infer the medical expenses spent by male, female smokers (1) and non-smokers (0). It is observed that, male smokers spend more than female smokers. At the same time, when comparing non-smoker gender and charges still male sex has the highest of medical expenses. This proves that male gender in USA are facing bad health issues when compared to female sex.

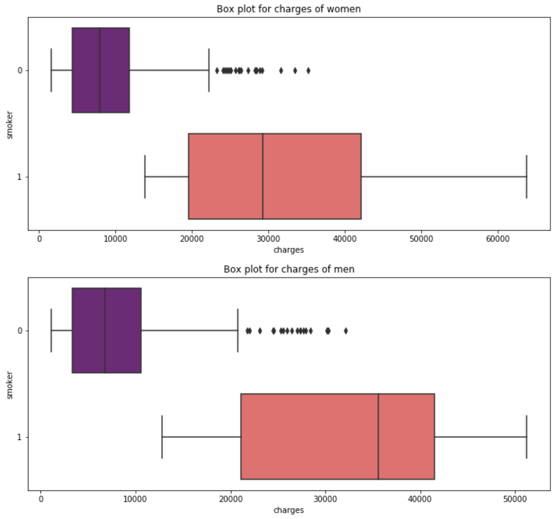


Fig 5: Box-plot for male and female expenditure

Box-plot for male and female (smokers/non-smokers) indicating the average expenditure of medical expenses. On an average, male smokers spend over 35000 dollars while female smokers are spending nearly 30000 dollars.

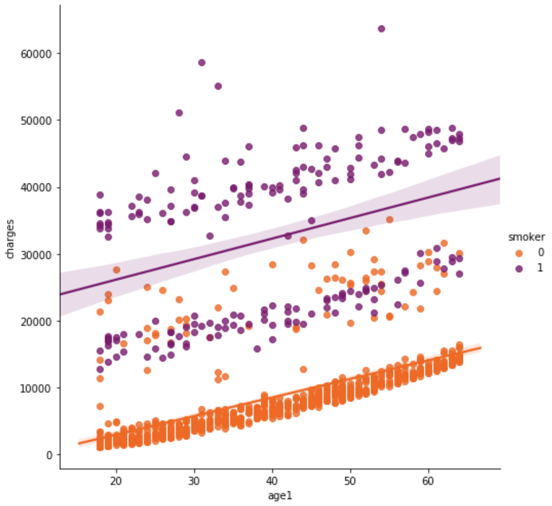


Fig 6: Smokers and non-smokers distribution based on age and charges

The above scatter plot is plotted against the age category and charges. This graph explains the medical expenses distribution from each age category. From the scatter plot we can observe linearity in data whereas age increases medical expenses also increases.

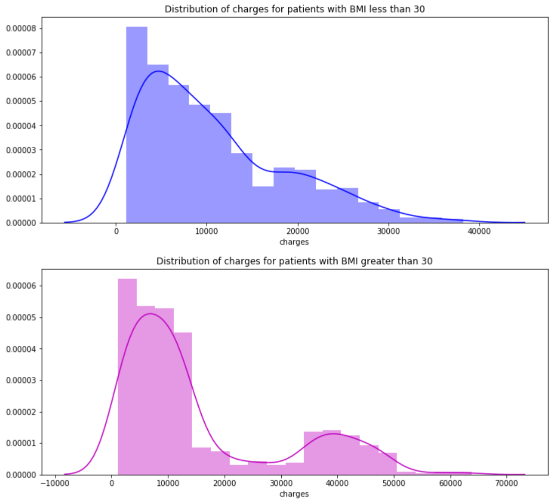


Fig 7: Medical expenses based on BMI

The above distribution graphs are plotted against BMI and medical expenses. Here, in our analysis average BMI is considered as 30 points where the BMI value greater than 30 starts obesity as per BMI chart. So, these plots are plotted against people with BMI < 30 and BMI > 30. The medical charges are compared between people who are obese and healthy. Obviously, the chart distribution was higher for people with BMI > 30 reaching almost 60000 dollars.

Average medical procedure cost

Average cost of expense on any procedure varies across different states in USA. The below chart plots the medical procedure of a cardiovascular disease and compares the cost between every state in the US. This shows a clear disparity in financial terms. This is just one of the thousands of procedures and drugs used in the country.

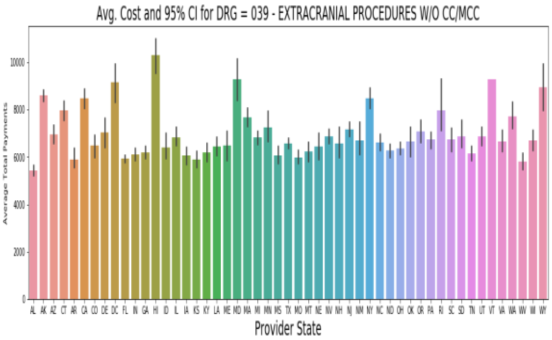


Fig 8: The figure shows the average cost of a medical procedure across all the states in USA.

In order to understand the implications of this in a more detailed manner, we decided that we need to go more deeper into the whole situation and decided to find out the cost of a procedure in each facility. For this we decided to use a swarm plot. The below graph gives a clear representation of the cost variation.

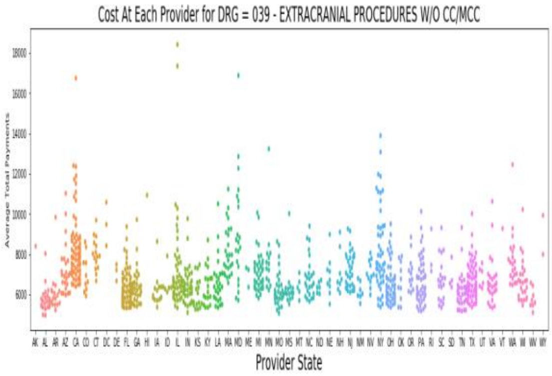


Fig 9: The above swarm plot for the cost of a procedure in various facilities of USA.

Data visualization based on Heart diseases Types of chest pain

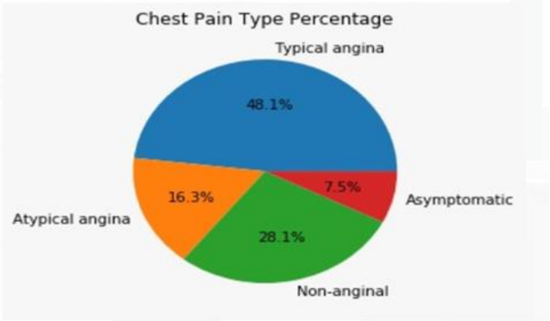


Fig 10: Pie-Chart showing rate of heart disease based of chest pain types

The Pie chart above we can see the percentage depending upon the types of chest pain. We can see that Typical angina has the highest percentage that is 48.1%. Then the second highest being the Non-anginal with 28.1%. After that comes the Atypical angina with a percent of 16.3% and the lowest being 7.5% that is Asymptomatic.

Chest Pain Type Vs Gender

In this bar graph we have separated the types of Chest pain based on Gender.

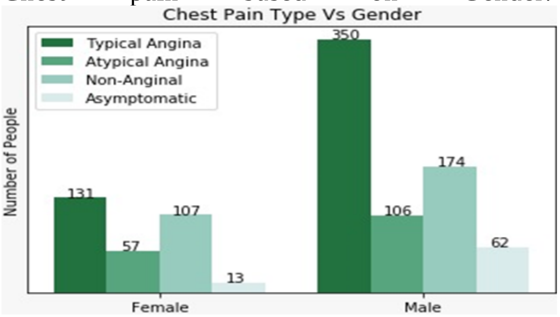


Fig 11- Male Vs Female chest pain ratio.

We can see that the Male show a higher number of people in all the types of chest pains compared to Female.

Heart health based on age group

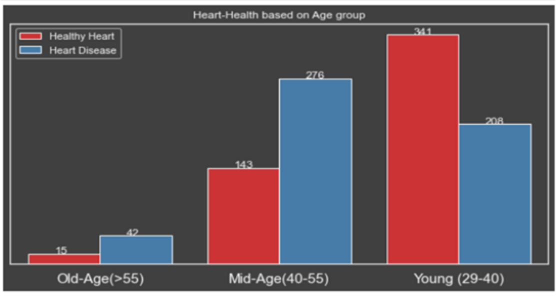


Fig 12- Bar-plot of heart health based on age.

Here is a plot of heart health based on age as we can see that the young age has the highest number of healthy hearts compared to all three types of age groups and the middle age group has the highest number of heart disease.

Heart health based on chest pain type

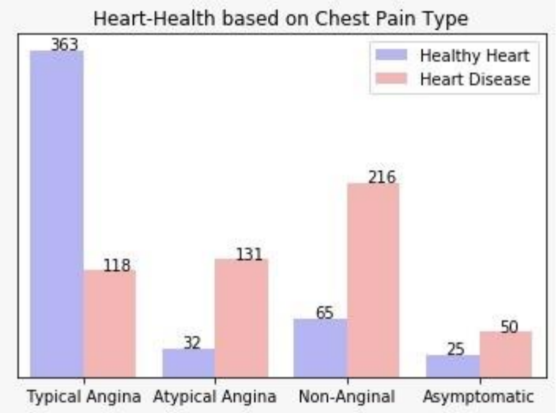


Fig 13 Types of chest pains.

Here is a bar graph which shows the health of the heart based on the chest pain we can see that the people having Typical Angina May not always have a diseased heart for the rest of the types there is a 50% or more chance to have the chest pain and having a diseased heart.

Implementation and Evaluation of Data mining models

Linear Regression model

Linear regression is applied when dependent variable is continuous and independent variable can be any form- discrete or continuous or indicator variable. The model is estimated by calculating one variable based on the value of another i.e. it quantifies the relationship between one independent variable and one outcome variable.

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Linear regression model was applied to predict the medical expenses of an individual. The model was able to produce an **accuracy of 65%** in predicting the medical charges of an individuals.

Random Forest

Random forest algorithm is used for solving expansive classification problems. It is a tree-based classifier, that uses multiple decision trees to build the model. It selects the maximum voted predictions as a result from the derived multiple decision trees. A random forest forms a strong classifier based on the collection of large number of weak classifiers.

We have also applied random forest model in predicting the medical expense. Surprisingly, the model preformed reasonably well in predicting medical charges with an **accuracy of 72%** gaining better result than linear regression model.

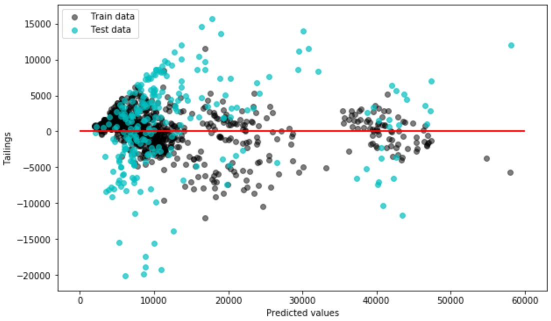
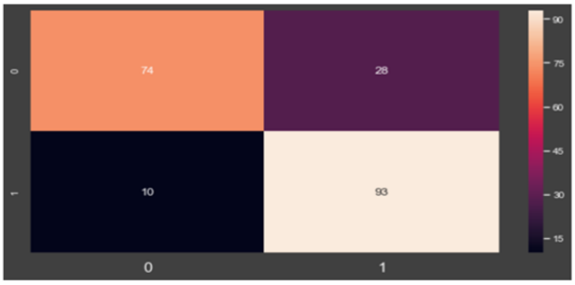


Fig 15: The scatter plot against the train data vs test data.

Logistic Regression

A classification algorithm that is applied to predict the categorical variables (values are 0's & 1's) based on the set of predictor variables. The model is used in this dataset 2 – Heart disease, since there is good linear dependency in data exists. Logistic regression was applied to classify and predict the "target" variable with disease or no disease. The model performed on a note in predicting the person disease with achieving outstanding **accuracy of 84%.**

Logistic regression Confusion matrix



VI. CONCLUSION

We were able to come up with a substantial set of analysis on the datasets that we chose and subsequently we were able to create interesting visualizations as well as applied statistical models to get a clearer idea.

For dataset 1, we can clearly see that **smokers spend 30000$** more than the average non-smoker. This was supported by the **random forest model** having the **accuracy of 72%** and **Linear Regression model** with **65% accuracy**.

For dataset 2, we can see that the different parameters for measuring the possibility of heart disease were highly accurate with the **Logistic regression** showing **84% of accuracy.**

For dataset 3, it is evident that the prices of different medical procedures or medicines vary by a big margin with the **CA state** topping in the highest average cost with **2000$**.

VII. REFERENCES

[1] S. Cohen, "The concentration of health care expenditures in the U.S. and predictions of future spending", Journal of Economic and Social Measurement, vol. 41, no. 2, pp. 167-189, 2016. Available: 10.3233/jem-160427 [Accessed 1 May 2020].

[2] K. Hsu, T. Yan and P. Liu, "A Study on the Annualized Medical Expense Prediction Model of the Bureau of National Healthy Insurance--The Application of the Grey Prediction Theory," 2006 IEEE International Conference on Systems, Man and Cybernetics, Taipei, 2006,

pp.764-769, doi: 10.1109/ICSMC.2006.384479.

[3] B. Sohrabi, I. Raeesi Vanani, N. Nikaein and S. Kakavand, "A predictive analytics of physicians prescription and pharmacies sales correlation using data mining", International Journal of Pharmaceutical and Healthcare Marketing, 2019. Available: 10.1108/ijphm-11-2017-0066 [Accessed 1 May 2020].

[4] A. Begum and A. Parkavi, "Prediction of thyroid Disease Using Data Mining Techniques," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp.342-345, doi: 10.1109/ICACCS.2019.8728320.

[5] S. Tithi, A. Aktar, F. Aleem and A. Chakrabarty, "ECG data analysis and heart disease prediction using machine learning algorithms", 2019 IEEE Region 10 Symposium (TENSYMP), 2019. Available: 10.1109/tensymp46218.2019.8971374 [Accessed 1 May 2020].

[6] E. Narbutas and L. Telksnys, "Data analysis for heart rate sequence elements in mobile systems", 2014 IEEE 2nd Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), 2014. Available: 10.1109/aieee.2014.7020323 [Accessed 1 May 2020].

[7] J. Katzenellenbogen et al., "Priorities for improved management of acute rheumatic fever and rheumatic heart disease: analysis of cross-sectional continuous quality improvement data in Aboriginal primary healthcare centres in Australia", Australian Health Review, vol. 44, no. 2, p. 212, 2020. Available: 10.1071/ah19132 [Accessed 1 May 2020].

[8] M. Jan, A. Awan, M. Khalid and S. Nisar, "Ensemble approach for developing a smart heart disease prediction system using classification algorithms", Research Reports in Clinical Cardiology, vol. 9, pp. 33-45, 2018. Available: 10.2147/rrcc.s172035 [Accessed 1 May 2020].

[9] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[10] D. Golinelli et al., "Health Expenditure and All-Cause Mortality in the ‘Galaxy’ of Italian Regional Healthcare Systems: A 15-Year Panel Data Analysis", Applied Health Economics and Health Policy, vol. 15, no. 6, pp. 773-783, 2017. Available: 10.1007/s40258-017-0342-x [Accessed 1 May 2020].