HEART DISEASE PREDICTION

A Project Work Synopsis

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

BIG DATA ANALYTICS

Submitted by:

DAKSH SHARMA

DILPREET KAUR

SHIV MURAT VERMA

IQBAL SINGH

University Roll Number

18BCS3772

18BCS3773

18BCS3768

18BCS3766

Under the Supervision of:

SUPERVISORS NAME

DR.OCHIN

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,
PUNJAB

MONTH & YEAR

NOVEMBER & 2020

Project Title and Brief Description

Predicting presence of Heart Diseases using Machine Learning

Machine Learning is used across many spheres around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

In this Synopsis, we'll discuss a project where I worked on predicting potential Heart Diseases in people using Machine Learning algorithms. The algorithms included K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier. The dataset has been taken from Kaggle.

Abstract

Machine learning (ML) is making a dramatic impact on cardiovascular magnetic resonance (CMR) in many ways. This review seeks to highlight the major areas in CMR where ML, and deep learning in particular, can assist clinicians and engineers in improving imaging efficiency, quality, image analysis and interpretation, as well as patient evaluation. We discuss recent developments in the field of ML relevant to CMR in the areas of image acquisition & reconstruction, image analysis, diagnostic evaluation and derivation of prognostic information. To date, the main impact of ML in CMR has been to significantly reduce the time required for image segmentation and analysis. Accurate and reproducible fully automated quantification of left and right ventricular mass and volume is now available in commercial products. Active research areas include reduction of image acquisition and reconstruction time, improving spatial and temporal resolution, and analysis of perfusion and myocardial mapping. Although large cohort studies are providing valuable data sets for ML training, care must be taken in extending applications to specific patient groups. Since ML algorithms can fail in unpredictable ways, it is important to mitigate this by open source publication of computational processes and datasets. Furthermore, controlled trials are needed to evaluate methods across multiple centers and patient groups.

Introduction

Machine learning (ML) and artificial intelligence (AI) are rapidly gaining importance in medicine, including in the field of medical imaging, and are likely to fundamentally transform clinical practice in the coming years. AI refers to the wider application of machines that perform tasks that are characteristic of human intelligence, e.g. infer conclusions from deduction or induction, while ML is a more restricted form of computational processing which uses a mathematical model together with training data to learn how to make predictions. Rather than explicitly computing results from a set of predefined rules, ML learns parameters from examples and therefore has the potential to perform better at a task such as detecting and differentiating patterns in data by being exposed to a more examples. The most advanced ML techniques, also called deep learning (DL), are especially well-suited for this purpose. Cardiovascular magnetic resonance (CMR) is a field that lends itself to ML because it relies on complex acquisition strategies, including multidimensional contrast mechanisms, as well as the need for accurate and reliable segmentation and quantification of biomarkers based on acquired data, to help guide diagnosis and therapy management.

Artificial intelligence (AI) can be seen as any technique that enables computers to perform tasks characteristic of human intelligence. Machine learning (ML) is generally seen as the subdiscipline of AI which uses a statistical model together with training data to learn how to make predictions. Deep learning (DL) is a specific form of ML that uses artificial neural networks with hidden layers to make predictions directly from datasets

It is important for clinicians and researchers working in CMR to understand the impact of ML on the field. Thus, the purpose of this review is threefold: firstly, we will provide a non-technical overview of the basics of ML relevant to CMR. Secondly, we survey the various ways ML has been applied to the field of CMR. Finally, we provide an outlook on future directions and recommendations for reporting results. Please also refer the glossary of terms for definitions of commonly used terms in machine learning.

Social Relevance, market need, Data collection and Literature Survey conducted

Social Relavence

The present study intends to help us understand, at a large scale, multivariate level, how the social determinants of health may influence cardiovascular diseases. What are some of the interrelationships among them and how do they fit into their particular geopolitical and historical contexts. To address these questions, we aim to take advantage of the vast corpus of literature already published in the field, while trying to remain as unbiased as possible in our analyses.

The approach we have decided to take implies resorting to a hybrid scheme involving a computational approach of automated literature mining and discourse analysis techniques supplemented with an ‘a posteriori’ discussion of the main findings.

In order to properly account for these phenomena, it is necessary to start with a consideration of those social determinants of health related to cardiovascular disease already identified –in the current literature– and then search for possible interrelationships among them. In this sense, the main purpose of this work is to perform a screening analysis based on a systematic search of the published scientific information about the SDCVD, and to use this analysis as a starting point to unveil their hidden complexities by resorting to data analytics, complex network analysis and visualization techniques that may allow us to see more clearly the relationship among them in order to better understand the effects of these social determinants on cardiovascular diseases [14].

In consequence, the approach followed in this study was as follows: First, we assembled a preliminary corpus by mining the entire PubMed database for all the papers related to social determinants of cardiovascular disease as denoted by corresponding Medical SubHeading (MeSH) classifiers. Then we performed a curation of the corpus to discard non-relevant content and code the information content, using both manual and automated techniques. Once we had a curated corpus, we built semantic networks (using co-occurrence of MeSH terms as links) and performed topological analyses of such networks to find associations between the different SDCVD. To provide context to such interrelationships, our study includes a detailed analysis of the historical trends. Finally, we discuss the main findings of all these stages aiming to contribute to generate an integrated framework for SDCVD

Market need

The cardiovascular disease market, which includes hypertension, dyslipidaemia and thrombotic events, is set to grow from $129.2 billion in 2015 to $146.4 billion by 2022, at a very modest compound annual growth rate of 1.8%, according to business intelligence provider GBI Research.

The company's latest report states that this relative stagnation can be attributed to major product approvals coinciding with key patent expirations. Within cardiovascular disease there are a number of blockbuster products that have recently gone off-patent, and others are expected to in the coming years, many of which belong to significant players.

For example, the current market leader, AstraZeneca's Crestor (rosuvastatin), generated around $7 billion in 2011, with revenues expected to drop sharply following the expiration of its patent on 8 July 2016. Total annual revenues are forecast to be around $1.3 billion in 2022.

Thomas Jarratt, associate analyst for GBI Research, explains: 'Unlike AstraZeneca, some key players will experience revenue growth resulting from the introduction of new products to market. In particular, Sanofi's Praluent (alirocumab) is expected to help mitigate losses associated with falling revenues of its key products Lovenox (enoxaparin) and Plavix (clopidogrel).

'Novartis' heart-failure drug Entresto was introduced to market in July 2015, and GBI Research expects its revenues to increase dramatically during the forecast period. Entresto is a combination drug, which has shown efficacy in clinical trials. Coupled with a high cost, which amounts to over $4 500 annually per patient, the drug contributes to a very high revenue forecast of $5.7 billion by 2022.'

The sheer number of expirations and approvals means the structure of the market will shift significantly. Current market leader AstraZeneca is set to mitigate the damage associated with the introduction of generic Crestor through the rising revenues attributed to its antiplatelet drug Brilinta.

Jarratt continues: 'the market shares of Sanofi and Novartis are expected to increase strongly over the forecast period, leading to Sanofi becoming market leader, and both brands achieving revenues in excess of $7 billion by 2022.'

Literature survey conducted

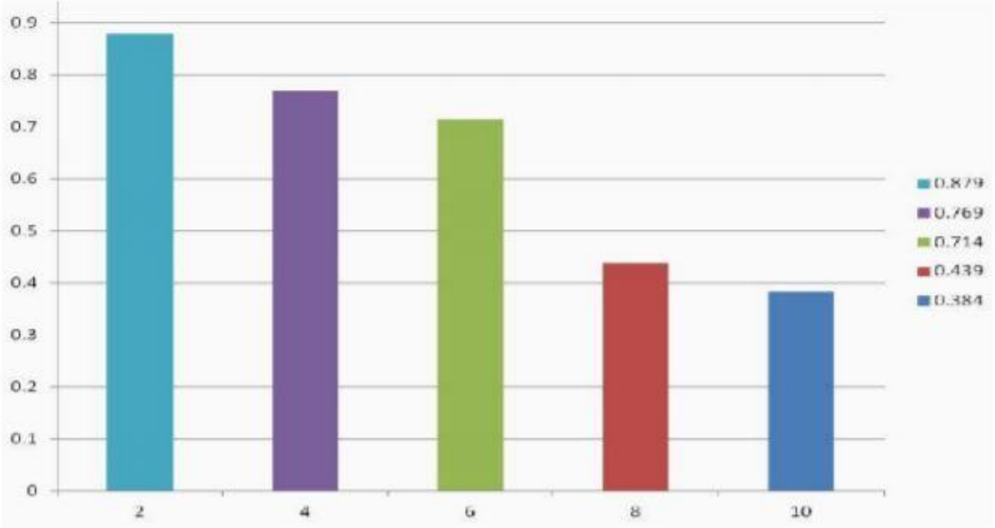
Till date different studies have been done on heart disease prediction. Various data mining and machine learning algorithms have been implemented and proposed on the datasets of heart patients and different results have been achieved for different techniques.

But, still today we are facing a lot of problem faced by the heart disease. Some of the recent research papers are as follows:

In 2010, A. Rajkumar and G. S. Reena applied machine learning algorithms such as Naïve Bayes, KNN (K- nearest neighbors) and decision list for heart disease prediction. Tanagra tool is used to classify the data and the data evaluated using 10-fold cross validation and the results are compared in table 4. The data set consists of 3000 instances with 14 different attributes. The dataset is divided into two parts, 70% of the data are used for training and 30% are used for testing. The results of comparison are based on 10-fold cross validation. Comparison is made among these classification algorithms out of which the Naive Bayes algorithm is considered as the better performance algorithm. Because it takes less time to build model and also gives best accuracy as compared to KNN and Decision Lists.

G.Subbalakshmi, K. Ramesh and M. Chinna Rao developed a Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modeling technique, namely, Naive Bayes. Using heart disease attributes such as chest pain, age, sex, cholesterol, blood pressure and blood sugar can predict the likelihood of patients getting a heart disease. It is implemented as web based questionnaire application. Historical data set of heart patients from Cleveland database of UCI repository was used to train and test the Decision Support System (DSS). The reasons to prefer Naive Bayes machine learning algorithm for predicting heart disease are as follows: when data is high, when the attributes are independent of each other and when we want to achieve high accuracy as compared to other models. When the dimensionality of the inputs is high in that case Naive Bayes classifier technique is particularly suited. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods

In 2011, M. A. Jabbar, Priti Chandra and B.L.Deekshatulu in this study develop a prediction system by implement associative rule mining using a new approach that combines the concept of sequence numbers and clustering for heart attract prediction. By using this approach first dataset of heart disease patients has been converted into binary format then apply proposed method on binary transitional data. Data set of heart disease patients has been taken from Cleveland database of UCI repository with 14 essential attributes. The algorithm is well known as Cluster Based Association Rule Mining Based on Sequence Number (CBARBSN). Support is a basic parameter in associative rule mining. To become element of a frequent item set an item should satisfy support threshold. In this research transactional data table is divided into clusters based on skipping fragments (disjoint sub sets of actual transitional table) then Sequence Number and Sequence ID of each item has been calculated. On the basis of Sequence ID frequent item sets has been discovered in different clusters and common frequent item set has taken as Global Item set. It has been observed from the experiment that Age>45 and Blood pressure>120 and Max Heart rate>100 and old Peak>0 and Thal>3 =>Heart attack (Common frequent item set found in both clusters in this experiment). In our proposed algorithm execution time to mine association rules is less (i.e., 0.879 ms when support=3) and as support increases execution time changes drastically as compared to previously developed system. In Fig Execution time is shown horizontally and Support vertically



Gap analysis — to validate and support your project idea

Typically, the primary goal of learning algorithms is to maximize the prediction accuracy or equivalently minimize the error rate. However, in the specific medical application problem we study, the ultimate goal is to alert and assist doctors in taking further actions to prevent hospitalizations before they occur, whenever possible. Thus, our models and results should be accessible and easily explainable to doctors and not only machine learning experts. With that in mind, we examine our models from two aspects: prediction accuracy and interpretability.

Prediction Accuracy
The prediction accuracy is captured in two metrics: the False Alarm Rate (the fraction of false positives out of the negatives) and the Detection Rate (the fraction of true positives out of the positives). Note that in the medical literature, the detection rate is often referred to as sensitivity and the term specificity is used for one minus the false alarm rate. For a binary classification system, the evaluation of the performance using these two metrics is typically illustrated with the Receiver Operating Characteristic (ROC) curve, which plots the Detection Rate versus the False Alarm Rate at various threshold settings.

Interpretability
With SVM, the features are mapped through a kernel function from the original space into a higher-dimensional space. This, however, makes the features in the new space not interpretable. In AdaBoost with trees, while a single tree classifier which is used as the base learner is explainable, the weighted sum of a large number of trees makes it relatively complicated to find the direct attribution of each feature to the final decision. The naïve Bayes Event model is in general interpretable, but in our specific problem each patient has a relatively small sequence of events (four) and each event is a composition of medical factors. Thus, again, to find the direct attribution of each feature to the final decision is hard. LRT itself and Logistic Regression still lack interpretability, because we have more than 200 features for each sample and there is no direct relationship between prediction of hospitalization and the reasons that led to it. The most interpretable method is K- LRT. K-LRT highlights the top K features that lead to the classification decision. These features could be of help in assisting the physicians reviewing the patient's EHR profile.

Defining Objectives of the projects

Import libraries

I imported several libraries for the project:

- numpy: To work with arrays
- pandas: To work with csv files and dataframes
- matplotlib: To create charts using pyplot, define parameters using rcParams and color them with cm.rainbow
- warnings: To ignore all warnings which might be showing up in the notebook due to past/future depreciation of a feature
- train_test_split: To split the dataset into training and testing data
- StandardScaler: To scale all the features, so that the Machine Learning model better adapts to the dataset

Next, I imported all the necessary Machine Learning algorithms

Import dataset

After downloading the dataset from Kaggle, I saved it to my working directory with the name dataset.csv. Next, I used read_csv() to read the dataset and save it to the dataset variable. Before any analysis, I just wanted to take a look at the data. So, I used the info() method.

As you can see from the output, there are a total of 13 features and 1 target variable. Also, there are no missing values so we don't need to take care of any null values. Next, I used describe() method.

The method revealed that the range of each variable is different. The maximum value of age is 77 but for chol it is 564. Thus, feature scaling must be performed on the dataset.

Understanding the data

Correlation Matrix

To begin with, let's see the correlation matrix of features and try to analyse it. The figure size is defined to 12 x 8 by using rcParams. Then, I used pyplot to show the correlation matrix. Using xticks and yticks, I've added names to the correlation matrix. colorbar() shows the colorbar for the matrix.

It's easy to see that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive.

Data Processing

To work with categorical variables, we should break each categorical column into dummy columns with 1s and 0s.

Let's say we have a column Gender, with values 1 for Male and 0 for Female. It needs to be converted into two columns with the value 1 where the column would be true and 0 where it will be false. Take a look at the Gist below.

To get this done, we use the get_dummies() method from pandas. Next, we need to scale the dataset for which we will use the StandardScaler. The fit_transform() method of the scaler scales the data and we update the columns.

Machine Learning

In this project, I took 4 algorithms and varied their various parameters and compared the final models. I split the dataset into 67% training data and 33% testing data.

K Neighbors Classifier

This classifier looks for the classes of K nearest neighbors of a given data point and based on the majority class, it assigns a class to this data point. However, the number of neighbors can be varied. I varied them from 1 to 20 neighbors and calculated the test score in each case.

Support Vector Classifier

This classifier aims at forming a hyperplane that can separate the classes as much as possible by adjusting the distance between the data points and the hyperplane. There are several kernels based on which the hyperplane is decided. I tried four kernels namely, linear, poly, rbf, and sigmoid.

Decision Tree Classifier

This classifier creates a decision tree based on which, it assigns the class values to each data point. Here, we can vary the maximum number of features to be considered while creating the model. I range features from 1 to 30 (the total features in the dataset after dummy columns were added).

Once we have the scores, we can then plot a line graph and see the effect of the number of features on the model scores.

Random Forest Classifier

This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features. Here, we can vary the number of trees that will be used to predict the class. I calculate test scores over 10, 100, 200, 500 and 1000 trees.

Next, I plot these scores across a bar graph to see which gave the best results. You may notice that I did not directly set the X values as the array [10, 100, 200, 500, 1000]. It will show a continuous plot from 10 to 1000, which would be impossible to decipher. So, to solve this issue, I first used the X values as [1, 2, 3, 4, 5]. Then, I renamed them using xticks.

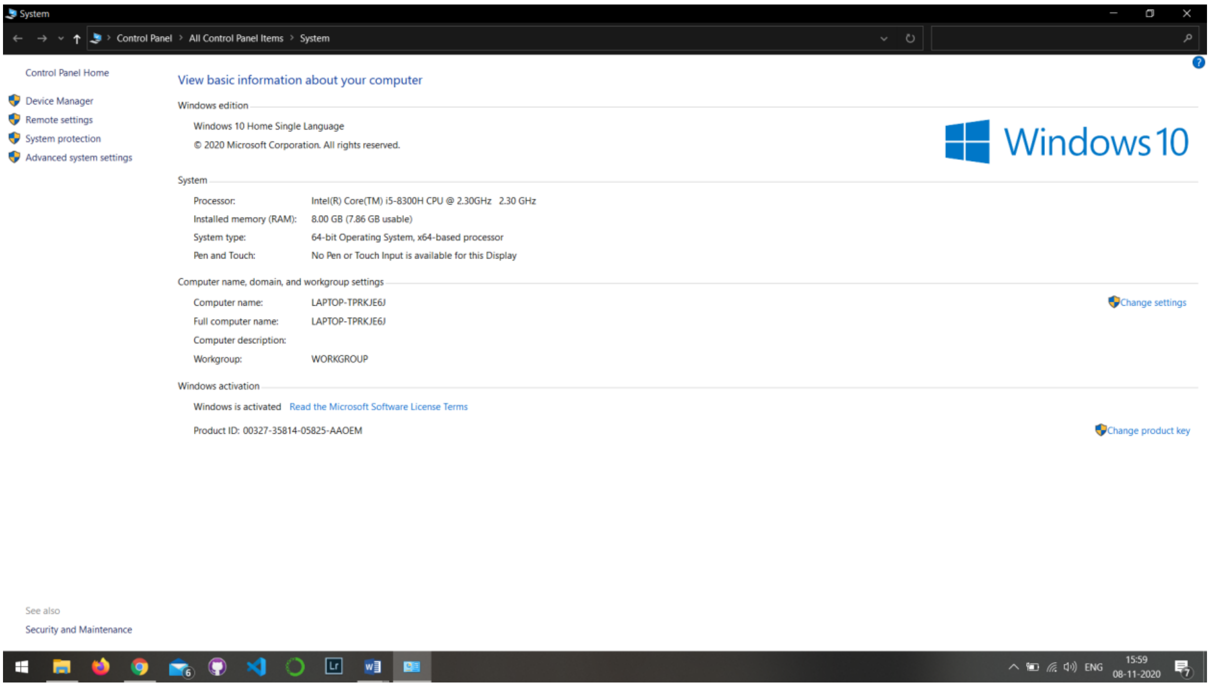
Conclusion

The project involved analysis of the heart disease patient dataset with proper data processing. Then, 4 models were trained and tested with maximum scores as follows:
K Neighbors Classifier: 87%
Support Vector Classifier: 83%
Decision Tree Classifier: 79%
Random Forest Classifier: 84%

Requirements and Constraints with respect to your project

Requirements

(hardware specifications)



View basic information about your computer

Windows edition

Windows 10 Home Single Language

© 2020 Microsoft Corporation. All rights reserved.

System

Processor: Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz 2.30 GHz

Installed memory (RAM): 8.00 GB (7.86 GB usable)

System type: 64-bit Operating System, x64-based processor

Pen and Touch: No Pen or Touch Input is available for this Display

Computer name, domain, and workgroup settings

Computer name: LAPTOP-TPRKJE6J

Full computer name: LAPTOP-TPRKJE6J

Computer description:

Workgroup: WORKGROUP

Windows activation

Windows is activated Read the Microsoft Software License Terms

Product ID: 00327-35814-05825-AAOEM

(software specification)

Visual Studio Code
Visual Studio Code is a source-code editor developed by Microsoft for Windows, Linux and macOS. It includes support for debugging, embedded Git control and GitHub, syntax highlighting, intelligent code completion, snippets, and code refactoring

Python 3
Python 3.0 (a.k.a. "Python 3000" or "Py3k") is a new version of the language that is incompatible with the 2.x line of releases. The language is mostly the same, but many details, especially how built-in objects like dictionaries and strings work, have changed considerably, and a lot of deprecated features have finally been removed. Also, the standard library has been reorganized in a few prominent places.

Anaconda
Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. Package versions are managed by the package management system conda.

Google colab
Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

Github
GitHub is a code hosting platform for version control and collaboration. It lets you and others work together on projects from anywhere. This tutorial teaches you GitHub essentials like repositories, branches, commits, and Pull Requests.

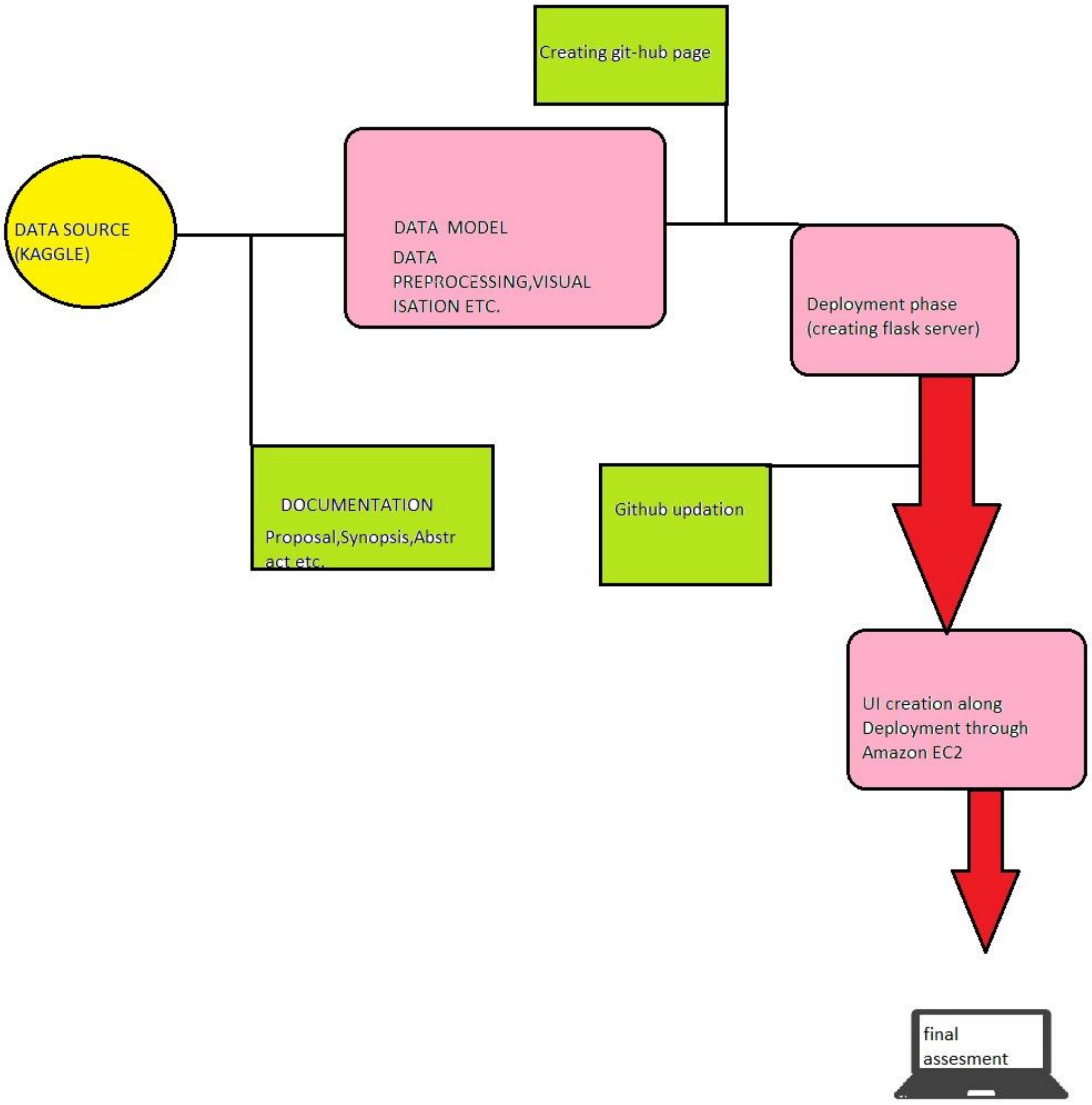
Constraints

it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

Machine learning (ML) proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry.

In this Project, We will be applying Machine Learning approaches(and eventually comparing them) for classifying whether a person is suffering from heart disease or not, using one of the most used dataset — Cleveland Heart Disease dataset from the UCI Repository.
As always, you can find the code used in this article in the Github Repository.

Design Specifications / Modules & Techniques along with Problem Formulation


Creating git-hub page
DATA SOURCE (KAGGLE)
DATA MODEL
DATA PREPROCESSING,VISUALISATION ETC.
Deployment phase (creating flask server)
DOCUMENTATION
Proposal,Synopsis,Abstract etc.
Github updation
UI creation along Deployment through Amazon EC2
final assesment

References

1. 1.

Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol. 2017;69:2657–64.

Article Google Scholar

2. 2.

Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol. 2018;71:2668–79.

Article Google Scholar

3. 3.

Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med. 2018;131:129–33.

Article Google Scholar

4. 4.

Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. Eur J Inter Med. 2018;48:e13–4.

CAS Article Google Scholar

5. 5.

Deo RC. Machine learning in medicine. Circulation. 2015;132:1920–30.

Article Google Scholar

6. 6.

Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35:1798–828.

Article Google Scholar

7. 7.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.

CAS Article Google Scholar

8. 8.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

Article Google Scholar

9. 9.

Oksuz I, Mukhopadhyay A, Dharmakumar R, Tsaftaris SA. Unsupervised myocardial segmentation for cardiac BOLD. IEEE Trans Med Imaging. 2017;36:2228–38.

Article Google Scholar

10.10.

Wang G, Zhang Y, Hegde SS, Bottomley PA. High-resolution and accelerated multi-parametric mapping with automated characterization of vessel disease using intravascular MRI. J Cardiovasc Magn Reson. 2017;19.

11. 11.

Baeßler B, Schaarschmidt F, Dick A, Stehning C, Schnackenburg B, Michels G, et al. Mapping tissue inhomogeneity in acute myocarditis: a novel analytical approach to quantitative myocardial edema imaging by T2-mapping. J Cardiovasc Magn Reson. 2015;17.

12.12.

Kramer CM, Barkhausen J, Flamm SD, Kim RJ, Nagel E. Standardized cardiovascular magnetic resonance imaging (CMR) protocols, society for cardiovascular magnetic resonance: board of trustees task force on standardized protocols. J Cardiovasc Magn Reson. 2008;10.

13.13.

Frick M, Paetsch I, den Harder C, Kouwenhoven M, Heese H, Dries S, et al. Fully automatic geometry planning for cardiac MR imaging and reproducibility of functional cardiac parameters. J Magn Reson Imaging. 2011;34:457–67.

Article Google Scholar

14.14.

Hayes C, Daniel D, Lu X, Jolly M-P, Schmidt M. Fully automatic planning of the long-axis views of the heart. J Cardiovasc Magn Reson. 2013;15.

15.15.

Goldfarb, JW, Cheng, J, Cao, JJ: Automatic optimal frequency adjustment for high field cardiac MR imaging via deep learning. In: CMR 2018 – A Joint EuroCMR/SCMR Meeting Abstract Supplement, pp. 437–438 (2018)