This is to certify that the project titled “------------------------------------’

is a bonafide work carried out by the following students in partial fulfillment of the requirements for

Advanced Academic Center intern, submitted to the chair, AAC during the academic year

\_\_2019-2020\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |
| --- | --- | --- |
| NAME | ROLL NO. | BRANCH |
| ANNE HEMANTH | 18241A04C1 | ECE |
| CHANDAVOLU SAI TEJA | 18241A04C4 | ECE |
| KARINGALA ARUSH | 18241A04D1 | ECE |
| GOPALDAS VAMSHI | 18241A04E0 | ECE |

This work was not submitted or published earlier for any study

Dr./Ms./Mr. Dr.B.R.K.Reddy Dr.Ramamurthy Suri

Program Coordinator Associate Dean, AAC

Project Supervisor

**ACKNOWLEDGEMENTS**

We express our deep sense of gratitude to our respected Director, Gokaraju Rangaraju Institute of Engineering and Technology for the valuable guidance and for permitting us to carry out this project.

With immense pleasure, we record our deep sense of gratitude tour respected principal, for permitting us to carry out this project.

We are thankful to Associate Dean, Advance Academic Centre for providing us appropriate ecosystem required for the project to complete.

We are thankful to our project supervisor who spared valuable time for us and influence novel ideas to guide us. I am indebted to all the above without whom I would not have concluded the project.

**Table of contents**

Authorization …………………………………………………………i

Acknowledgments………………………………………….…………ii

Abstract……………………………………...................…………….iii

1. Introduction………………………………………………………1-4

* 1. What is ML……………………………………………….
  2. Data Pre-Processing………………………………….……
  3. What is classification algorithm………………………….
  4. Logistic regression………………………………………..

1. Data Set……………………………………………………….

3. Logistic Regression Algorithm……………………………….

4. Result

4.1Output

**Abstract**

Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, accurate detection of heart diseases in all cases and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise.

In this study, a tentative design of a ML-based heart disease prediction system had been proposed to detect impending heart disease using Machine learning techniques. For the accurate detection of the heart disease, an efficient machine learning technique should be used which had been derived from a distinctive analysis among several machine learning algorithms in a Python Platform. The proposed algorithm is a logistic regression technique which is a classification algorithm, by this using the parameters one can predict that if a person will be effected by a heart attack or not.

**Introduction**

The heart is a kind of muscular organ which pumps blood into the body and is the central part of the body’s cardiovascular system which also contains lungs.

Cardiovascular system also comprises a network of blood vessels, for example,veins, arteries, and capillaries. These blood vessels deliver blood all over the body. Abnormalities in normal blood flow from the heart cause several types of heart diseases which are commonly known as cardiovascular diseases (CVD).

Heart diseases are the main reasons for death worldwide. According to the survey of the World Health Organization (WHO), 17.5 million total global deathsoccur because of heart attacks and strokes. More than 75% of deaths from cardiovascular diseases occur mostly in middle-income and low-income countries.

Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack. Therefore, detection of cardiac abnormalities at the early stage and tools for the prediction of heart diseases can save a lot of life and help doctors to design an effective treatment plan which ultimately reduces the mortality rate due to cardiovascular diseases. Due to the development of advance health care systems, lots of patient data are nowadays available (i.e. Big Data in Electronic Health Record System) which can be used for designing predictive models for Cardiovascular diseases. Data mining or machine learning is a discovery method for analyzing big data from an assorted perspective and encapsulating it into useful information. “Data Mining is a non-trivial extraction of implicit, previously unknown and potentially useful information about data” . Nowadays,a huge amount of data pertaining to disease diagnosis, patients etc. are generated by health care industries. Machine Learning provides a number of techniques which

discover hidden patterns or similarities from data. Therefore, in this paper, a machine learning algorithm is proposed for the implementation of a heart dis-ease prediction system which was validated on two open access heart disease prediction datasets.

Machine learning in medicine has recently made headlines. [Google has developed a machine learning algorithm](https://www.mercurynews.com/2017/03/03/google-computers-trained-to-detect-cancer/) to help identify cancerous tumors on mammograms. [Stanford is using a deep learning algorithm](https://news.stanford.edu/2017/01/25/artificial-intelligence-used-identify-skin-cancer/)to identify skin cancer.Machine learning in medicine has recently made headlines. [Google has developed a machine learning algorithm](https://www.mercurynews.com/2017/03/03/google-computers-trained-to-detect-cancer/) to help identify cancerous tumors on mammograms. [Stanford is using a deep learning algorithm](https://news.stanford.edu/2017/01/25/artificial-intelligence-used-identify-skin-cancer/)to identify skin cancer.

**What is Machine Learning**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

### **Some machine learning methods**

Machine learning algorithms are often categorized as supervised or unsupervised.

* ****Supervised machine learning algorithms****can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
* In contrast, ****unsupervised machine learning algorithms****are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
* ****Semi-supervised machine learning algorithms**** fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn’t require additional resources.
* ****Reinforcement machine learning algorithms****is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

**Data Preprocessing**

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc.

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

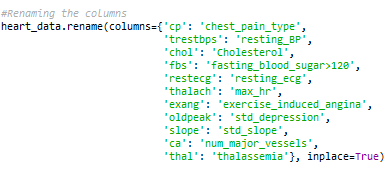
The data must be pre-processed before feeding and fitting it to the algorithm. For the given Dataset we need to rename the table headers.

There is no missing data or categorical data to pre-process and logistic regression does not require feature scaling as this is done by the algorithm itself.

**Renaming the table headers**

The given dataset has titles(table headers) in short form, this needs to be converted into formal and actual form so that we can easily find the relations between the independent attributes.

This can be done by the below code snippet in python.



**Classification Algorithms**

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

Here we have the types of classification algorithms in Machine Learning:

1. Linear Classifiers: Logistic Regression, Naive Bayes Classifier
2. Nearest Neighbor
3. Support Vector Machines
4. Decision Trees
5. Boosted Trees
6. Random Forest
7. Neural Networks

## **Naive Bayes Classifier (Generative Learning Model) :**

It is a classification technique based on [Bayes’ Theorem](https://en.wikipedia.org/wiki/Bayes'_theorem?source=post_page---------------------------) with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

## **Nearest Neighbor:**

The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the “k” is the number of neighbors it checks).

## ****Logistic Regression (Predictive Learning Model) :****

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. This is better than other binary classification like nearest neighbor since it also explains quantitatively the factors that lead to classification.

## ****Decision Trees:****

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

## ****Random Forest:****

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees’ habit of over fitting to their training set.

## ****Neural Network:****

A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. Generally the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand.

**Logistic Regression**

In [statistics](https://en.wikipedia.org/wiki/Statistics" \o "Statistics), the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be combined to model several classes of events such as determining whether an image contains a cat, dog, lion, etc... Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

Logistic regression is a [statistical model](https://en.wikipedia.org/wiki/Statistical_model" \o "Statistical model) that in its basic form uses a [logistic function](https://en.wikipedia.org/wiki/Logistic_function" \o "Logistic function) to model a [binary](https://en.wikipedia.org/wiki/Binary_variable" \o "Binary variable) [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable" \o "Dependent variable), although many more complex [extensions](https://en.wikipedia.org/wiki/Logistic_regression" \l "Extensions) exist. In [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis" \o "Regression analysis), logistic regression (or logit regression) is [estimating](https://en.wikipedia.org/wiki/Estimation_theory" \o "Estimation theory) the parameters of a logistic model (a form of [binary regression](https://en.wikipedia.org/wiki/Binary_regression" \o "Binary regression)). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an [indicator variable](https://en.wikipedia.org/wiki/Indicator_variable" \o "Indicator variable), where the two values are labeled "0" and "1". In the logistic model, the [log-odds](https://en.wikipedia.org/wiki/Log-odds" \o "Log-odds) (the [logarithm](https://en.wikipedia.org/wiki/Logarithm" \o "Logarithm) of the [odds](https://en.wikipedia.org/wiki/Odds" \o "Odds)) for the value labeled "1" is a [linear combination](https://en.wikipedia.org/wiki/Linear_function_(calculus)" \o "Linear function (calculus)) of one or more [independent variables](https://en.wikipedia.org/wiki/Independent_variable" \o "Independent variable) ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a [continuous variable](https://en.wikipedia.org/wiki/Continuous_variable" \o "Continuous variable) (any real value). The corresponding [probability](https://en.wikipedia.org/wiki/Probability" \o "Probability) of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The [unit of measurement](https://en.wikipedia.org/wiki/Unit_of_measurement" \o "Unit of measurement) for the log-odds scale is called a *[logit](https://en.wikipedia.org/wiki/Logit" \o "Logit)*, from *logistic unit*, hence the alternative names. Analogous models with a different [sigmoid function](https://en.wikipedia.org/wiki/Sigmoid_function" \o "Sigmoid function) instead of the logistic function can also be used, such as the [probit model](https://en.wikipedia.org/wiki/Probit_model" \o "Probit model); the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the [odds ratio](https://en.wikipedia.org/wiki/Odds_ratio" \o "Odds ratio).

Logistic regression can be binomial, ordinal or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable" \o "Dependent variable) can have only two possible types, "0" and "1" (which may represent, for example, "dead" vs. "alive" or "win" vs. "loss"). [Multinomial logistic regression](https://en.wikipedia.org/wiki/Multinomial_logit" \o "Multinomial logit) deals with situations where the outcome can have three or more possible types (e.g., "disease A" vs. "disease B" vs. "disease C") that are not ordered. [Ordinal logistic regression](https://en.wikipedia.org/wiki/Ordinal_logistic_regression" \o "Ordinal logistic regression) deals with dependent variables that are ordered.

In binary logistic regression, the outcome is usually coded as "0" or "1", as this leads to the most straightforward interpretation.[[14]](https://en.wikipedia.org/wiki/Logistic_regression" \l "cite_note-Hosmer-14) If a particular observed outcome for the dependent variable is the noteworthy possible outcome (referred to as a "success" or a "case") it is usually coded as "1" and the contrary outcome (referred to as a "failure" or a "noncase") as "0". Binary logistic regression is used to predict the [odds](https://en.wikipedia.org/wiki/Odds" \o "Odds) of being a case based on the values of the [independent variables](https://en.wikipedia.org/wiki/Independent_variable" \o "Independent variable) (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a noncase.

**DATASET**

This database contains 14 attributes . In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

### Content **Attribute Information:**  > 1. age  > 2. sex  > 3. chest pain type (4 values)  > 4. resting blood pressure  > 5. serum cholestoral in mg/dl  > 6. fasting blood sugar > 120 mg/dl > 7. resting electrocardiographic results (values 0,1,2) > 8. maximum heart rate achieved  > 9. exercise induced angina  > 10. oldpeak = ST depression induced by exercise relative to rest  > 11. the slope of the peak exercise ST segment  > 12. number of major vessels (0-3) colored by flourosopy  > 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect.

### Acknowledgements

### Creators:  1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.  2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.  3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.  4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

### Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

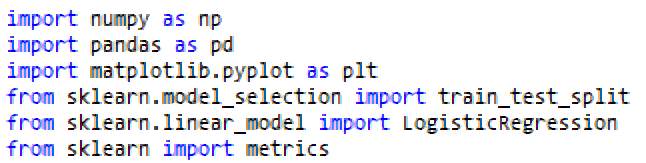
**Logistic Regression Algorithm**

The algorithm is based on the concept of logistic function to model a binary dependent variable.

The algorithm is as follows

**Importing the necessary libraries**

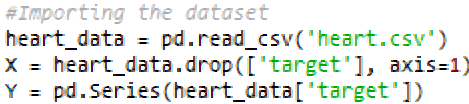
The entire concept of Machine Learning is built upon the concept of Object Oriented Programming hence we need to import the necessary packages before starting the hardcode the program. This list of necessary packages and the python code to import these packages is given in the below code snippet.



**Importing the Dataset**

The dataset should be imported to the algorithm in order to fit the classifier to the dataset, and the dataset should be divided into the set of independent variables and dependent variables. So that the algorithm will infer on the relation between each column(independent variable) and the dependent variable.

The code is given in the following code snippet.



Y is the target attribute or the dependent variable, the title of the column is target hence we use python code to pick up the attribute with the title 'target'.

X is the set of independent attributes or the attributes that actually give the relationship to the heart health of a specific person. We drop the 'target' attribute and pickup all other attributes to make this X dataframe.

**Renaming the titles**

Refer to Data Preprocessing Section

**Spliting the dataset**

The general purpose of a machine learning algorithm is to learn on some data and then apply this knowledge on new data elements to predict the future possibilities of a scenario. Hence we need to check the accuracy of the algorithm hence we train our algorithm on the training set and then we test its accuracy on the testing set. Hence we divide the dataset into 70% training and 20% testing sets.

The code snippet for doing this is given below.

Screenshot (172)

**Fitting the Algorithm**

As the necessary package was initially imported. The function needs to be called and the algorithm should be applied on the training data that is the X\_train and Y\_train, using this code the algorithm we fit on the given dataset and will infer on the data and acquire knowledge based on the relations that occur between the independent and the dependent data.

Screenshot (175)

**Predicting on testing set.**

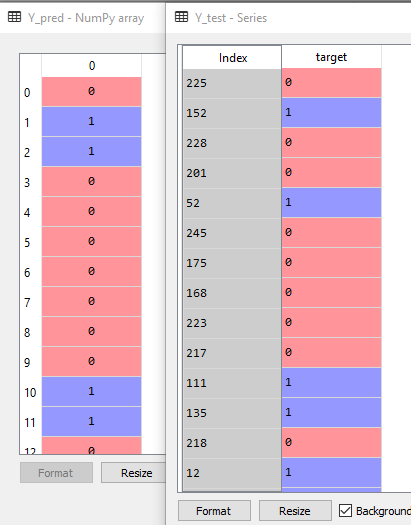
Once the algorithm is fitted on the training dataset and is trained . It has accquired the knowledge about the relations between the independent variables and the dependent variable.

Now it has to be tested on new data that it was never trained on to compare the accuracy of the algorithm based on the this new testing data. For this a prediction table is generated based on the predicted values generated on the testing data i.e X\_test and this generated data can be called as Y\_pred or (predicted data).

The code snippet for this piece of code is given below.

Screenshot (176)

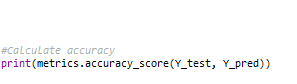
This will generate data predicted values of the testing dataset. Now in this case we have Y\_test which is the actual true value from the data set and Y\_pred which is the predicted value. Hence these are comparable to know the accuracy of the algorithm.



**Accuracy Check**

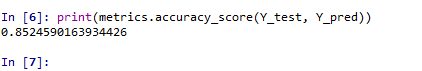
To check the accuracy of the algorithm that we have built we can use a code snippet for comparing the number of 1's and the number of 0's for each of the actual case for which it has predicted correctly and for which it has predicted incorrectly.

The below code snippet has the code necessary for that.



Accuracy

Hence the given algorithm has shown 85% Accuracy.



**RESULTS**

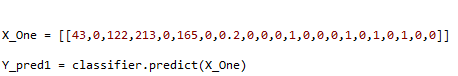
The result of this Machine Learning Classifier is to predict weather a person with specific parameters of given health conditions is vulnerable of getting a heart attack or not.

Hence with a accuracy of 85% we can predict weather a person is vulnerable of getting a heart attack or not.

**OUTPUT**

For a given set of attributes it is checked weather the person is vulnerable to heart attack or not.

A code snippet is given below



The output is given below, which correlates perfectly 