**JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY**



Exploration and evaluation of prediction accuracy of various machine learning algorithms on heart disease data set .

**MINOR PROJECT REPORT**

**EVEN SEMESTER 2018**

|  |  |  |
| --- | --- | --- |
| **AKASH KUMAR** | **15103168** | **B3** |
| **SAURABH GUPTA** | **15103193** | **B3** |
| **KRISHNA NAYAK** | **15103091** | **B3** |

**Submitted To:**

**DR. P. RAGHU WAMSI**

**OBJECTIVE**

The objective of the paper is to explore and evaluate the heart disease prediction accuracy of various machine learning algorithms on heart disease data set.

The examined machine learning algorithms are:

1. Simple Linear regression
2. Multiple linear regression
3. Logistic regression
4. Decision Tree
5. Support Vector machine
6. K means clustering

**ABSTRACT**

Everyone is a patient at some time or another, and we all want good medical care. We assume that doctors are all medical experts and that there is good research behind all their decisions. However, that cannot always be the case. Nevertheless, they cannot possibly commit to memory all the knowledge they need for every situation, and they probably do not have it readily available. Even if they did have access to the massive amounts of data needed to compare treatment outcomes for all the diseases they encounter, they would still need time and expertise to analyze that information and integrate it with the patient's own medical profile. In this era of advancements and every service on your finger tips only by using your attractive gadgets like Mobile phone and PC etc. So Now we can shop online and can sell online. We can study online, chat online even telecommunication in the form of a videoconference. We can read news ,find jobs, and work from home in A/C with PC. The most accurate prediction of whats going inside your body, without getting too much inside your body and without , can be made from your heart beat data and now a days there are various watches that keep a watch on your heart beat all day long .

So we want to process the heart beat data and make various prediction what all are going inside your body and whether everything inside is right or not and is your body under pressure or any critical condition that needs to be taken care of , based on machine learning approach.

**INTRODUCTION**

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It’s a science that’s not new – but one that has gained fresh momentum.

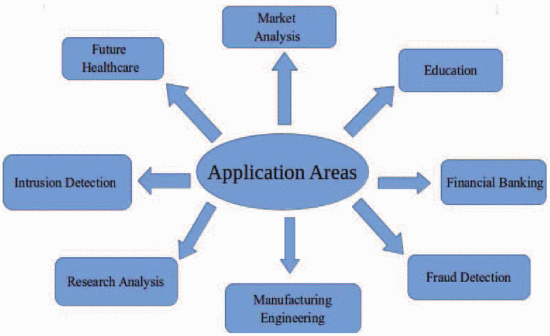
While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster – is a recent development. Here are a few widely publicized examples of machine learning applications you may be familiar with:

* The heavily hyped, self-driving Google car? The essence of machine learning.
* Online recommendation offers such as those from Amazon and Netflix? Machine learning applications for everyday life.
* Knowing what customers are saying about you on Twitter? Machine learning combined with linguistic rule creation.
* Fraud detection? One of the more obvious, important uses in our world today.

**TECHNOLOGY USED**

### 1. Machine Learning

We are entering era of big data. For example, there are about 1 trillion web pages; 1 hour of video is uploaded to youTube each second amounting to ten years of content every day.So this overflow of data calls for preset methods of data analysis which is what machine learning provides. We define ML as a set of methods that can automatically identify patterns in data and then use the uncovered patterns to forecast future data or to perform other kinds of decision making under uncertainty. ML is a set of tools that rarely speaking allows us to “teach” computers how to perform tasks by providing the examples of how they should be done.

* 

2.**PYTHON PROGRAMMING**

Python is a multiparadigm, general-purpose, interpreted, high-level programming language. Python allows programmers to use different programming styles to create simple or complex programs, get quicker results and write code almost as if speaking in a human language.

**NOVELTY OF PROBLEM**

Healthcare is arguably one of the most complex, challenging and expensive industries there is. Providers and professionals are nearly always on the back foot, managing diseases and health conditions reactively, often without all the information they need to provide the most appropriate care. This is the problem is solving, using machine learning to predict not only who will get sick, but also how sick they’ll get and helping coordinate their care more effectively. It does this by pulling data from a range of sources including patient devices, electronic medical and public records to identify patterns and predict future health. It’s already being used across 11 health systems in the US to treat diseases including cancer, sepsis and heart attacks. The business has also just raised $8.5m in its Series A investment round.The challenge in healthcare today is slightly different from robotics. It is more basic. It has to do with how the healthcare industry is dealing with data. Healthcare companies are attacking data sequentially – which is exactly what they should be doing. Healthcare execs will tell you firsthand that, as an industry, they are behind the financial services, technology and energy sectors when it comes to maximizing data as a resource. This step is particularly challenging due to both the volume of the data as well as the dimensionality of that data. Standard approaches, those used over the past decade or so, break down when confronted with either volume or dimensionality/complexity much less both.

**UNIQUENESS OF THE IMPLEMENTATION**

When a physician arrives at a diagnosis, it is the result of years of learning, their personal experience with similar patients and being up to date with the latest developments in the field. This is a huge amount of information to retrieve and analyze at all times, which doctors do with admirable success.

But the reality is that the mental capacity of an individual is limited and one physician can only treat that many patients in their life.

Machine learning has been used to augment the physician’s capacity by looking at the total available data on the patient and making recommendations based on this information - This approach is only limited by the total number of patients treated in the world.Machine learning could be used here to sift through the data and present the doctor with the information they need to make a decision.

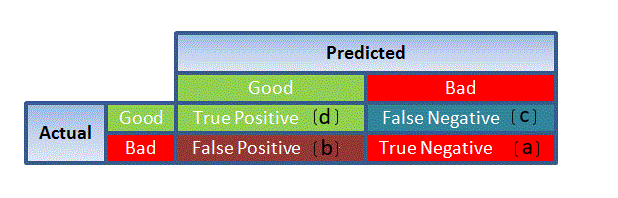
A machine learning approach to test the heart beat data and process it under different machine learning algorithms and determine a best algorithm which can handle the data and give the most accurate prediction of the processed data.

Also,

-> Pros and Cons of each algorithms

* -> Which algorithm can handle these continuous data
* set and process it best.
* -> How the implimentation of each algorithms make various prediction.
* -> What all each algorithm is capable of predicting .
* -> And what is the accuracy of each predition through various algorithms.

**IMPLEMENTATION**

* **Linear Regression model:** Simple linear regression is a statistical method that enables users to summarise and study relationships between two continuous (quantitative) variables. Linear regression is a linear model wherein a model that assumes a linear relationship between the input variables (x) and the single output variable (y). Here the y can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is called a simple linear regression. A linear regression line has an equation of the form ***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).
* **Multiple linear Regression:** When there are multiple input variables, the procedure is referred as multiple linear regression.
* the model for multiple linear regression, given *n* observations, is   
  *y*i = http://www.stat.yale.edu/Courses/1997-98/101/beta.gif0 + http://www.stat.yale.edu/Courses/1997-98/101/beta.gif1*x*i1 + http://www.stat.yale.edu/Courses/1997-98/101/beta.gif2*x*i2 + ... http://www.stat.yale.edu/Courses/1997-98/101/beta.gifp*x*ip + http://www.stat.yale.edu/Courses/1997-98/101/eps.gifi for *i* = 1,2, ... *n*.
* **Logistic Regression:** Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). In 1972, Nelder and Wedderburn proposed this model with an effort to provide a means of using linear regression to the problems which were not directly suited for application of linear regression. Infact, they proposed a class of different models (linear regression, ANOVA, Poisson Regression etc) which included logistic regression as a special case.The fundamental equation of generalized linear model is:
* g(E(y)) = α + βx1 + γx2
* **Confusion Matrix:** It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like:
* [](https://www.analyticsvidhya.com/wp-content/uploads/2015/11/1111.png)
* **DECISION TREE:** A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called “root” that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values.

**RELATED WORK**

* In the paper of B. K. Bhavitha ,Anisha P. Rodrigues and Nirnan Chiplunkar presented at International conference on inventive Communication and computational technologies(ICICCT,2017)includes outline of current works that done on sentimental classification and analysis. From the survey we can conclude that supervised learning methods like Naive Bayesian and Support Vector Machine are considered as standard learning method. Support Vector Machine provides excellent accuracy as compared to many other classifiers. In terms of accuracy we concluded that with small feature set Naive Bayes performs well, if large feature set is taken then SVM will be the best choice.
* In The 2010 International Joint Conference on Neural Networks (IJCNN) we can say that according to the results of the machine learning techniques that we compare in this work, that we have a good performance for the spam filtering in blog comments particular case. We can mention that the support vector machine technique has a better performance that the Naïve Bayes and the k-Nearest Neighbors techniques and a slightly better performance with respect to the neural networks technique.
* In the research paper of Amanpreet Singh, Narina Thakur, Aakansha Sharma review of different machine learning algorithms stated  a comprehensive review of the key ideas, drawing out pros and cons and useful variants of the discussed algorithms. The paper shows that every algorithm differs according to area of application and it is not the case that a single algorithm is superior in every scenario. The decision of choosing an appropriate algorithm is based on the type of problem and the data available. Again, by choosing two or more suitable algorithm and creating an ensemble, the accuracy can be increased.

**PRELIMINARIES AND ASSUMPTIONS**

What's required to create good machine learning systems?

* Data preparation capabilities.
* Algorithms – basic and advanced.
* Automation and iterative processes.
* Scalability.
* Ensemble modeling.

We have implemented the functionalities

**DIFFERENT MACHINE LEARNING TECHNIQUES**

## What are some popular machine learning methods?

Two of the most widely adopted machine learning methods are **supervised learning** and **unsupervised learning** – but there are also other methods of machine learning. Here's an overview of the most popular types.

**Supervised learning**algorithms are trained using labeled examples, such as an input where the desired output is known. For example, a piece of equipment could have data points labeled either “F” (failed) or “R” (runs). The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data. Supervised learning is commonly used in applications where historical data predicts likely future events. For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim.

**Unsupervised learning**is used against data that has no historical labels. The system is not told the "right answer." The algorithm must figure out what is being shown. The goal is to explore the data and find some structure within. Unsupervised learning works well on transactional data. For example, it can identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find the main attributes that separate customer segments from each other. Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition. These algorithms are also used to segment text topics, recommend items and identify data outliers.

**Semisupervised learning**is used for the same applications as supervised learning. But it uses both labeled and unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data (because unlabeled data is less expensive and takes less effort to acquire). This type of learning can be used with methods such as classification, regression and prediction. Semisupervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process. Early examples of this include identifying a person's face on a web cam.

**Reinforcement learning** is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. This type of learning has three primary components: the agent (the learner or decision maker), the environment (everything the agent interacts with) and actions (what the agent can do). The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good policy. So the goal in reinforcement learning is to learn the best policy.

### Who's using it?

#### Most industries working with large amounts of data have recognized the value of machine learning technology. By gleaning insights from this data – often in real time – organizations are able to work more efficiently or gain an advantage over competitors.

### Financial services

Banks and other businesses in the financial industry use machine learning technology for two key purposes: to identify important insights in data, and prevent fraud. The insights can identify investment opportunities, or help investors know when to trade. Data mining can also identify clients with high-risk profiles, or use cyber surveillance to pinpoint warning signs of fraud.

### Government

Government agencies such as public safety and utilities have a particular need for machine learning since they have multiple sources of data that can be mined for insights. Analyzing sensor data, for example, identifies ways to increase efficiency and save money. Machine learning can also help detect fraud and minimize identity theft.

### Health care

Machine learning is a fast-growing trend in the health care industry, thanks to the advent of wearable devices and sensors that can use data to assess a patient's health in real time. The technology can also help medical experts analyze data to identify trends or red flags that may lead to improved diagnoses and treatment.

### Marketing and sales

Websites recommending items you might like based on previous purchases are using machine learning to analyze your buying history – and promote other items you'd be interested in. This ability to capture data, analyze it and use it to personalize a shopping experience (or implement a marketing campaign) is the future of retail.

### Oil and gas

Finding new energy sources. Analyzing minerals in the ground. Predicting refinery sensor failure. Streamlining oil distribution to make it more efficient and cost-effective. The number of machine learning use cases for this industry is vast – and still expanding.

### Transportation

Analyzing data to identify patterns and trends is key to the transportation industry, which relies on making routes more efficient and predicting potential problems to increase profitability. The data analysis and modeling aspects of machine learning are important tools to delivery companies, public transportation and other transportation organizations.

**STEPS AND SIMULATION**

Step1: we have implemented numpy library for numerical python arrays etc., pandas library for data base , matplotlib and matplotlib.pyplot for graphs and plotting relations.

Step2:We imported benchmark dataset.

Step3:For each implemented algorithm , we have used splitting the data into training and test data sets

Step3: we have done featured scaling for different attributes.

Step4: Fitting algorithms to training datasets.

Step5: Prediction of results.

Step6: We have created the confusion matices for each algorithm.

Step7: encoding of categorical data where required.

Step8: LabelEncoder and Onehotencoder.

Step9: building the optimal model using backward elimination.

Comparison between machine learning methods:

|  |  |  |
| --- | --- | --- |
| **ALGORITHM** | **ADVANTAGE** | **DISADVANTAGE** |
| Simple Linear Regression | Very simple to implement  Find the relation of two variables | Only dependent and independent variable relation  Incorrect Straight Line |
| Multiple Linear Regression | Ability to determine the relative influence of predictor variables to the criterion value.  Ability to identify outliers and anomalies | Pitfalls of incomplete data. |
| Logistic Regression | It is more robust.  The variable don’t need to normally distributed. | No homogeneity of various assumptions.  Normally distributed terms are not assumed. |
| Decision Tree | This is very fast in learning data set.  Easy for understanding purpose. | It has problem of difficulty to handle data with noisy data.  Over fitting of data. |
| Support Vector Machine | High accuracy even with large data set .  Works well with many number of dimensions  No over fitting | Problem in representing document into numerical vector |
| K means algo | Computationally faster for large dataset.  Produce tighter clusters | Difficult to predict K-value.  With global cluster it did not work well. |
| Naive Bayes algo | Simple and work well with textual as well as numerical data  Easy to implement  Computationally cheap | Perform very poorly when feature set is highly correlated  It gives relatively low classification performance for large data set.  Independent assumption of attributes may lead to inaccurate result |

**GRAPHS**

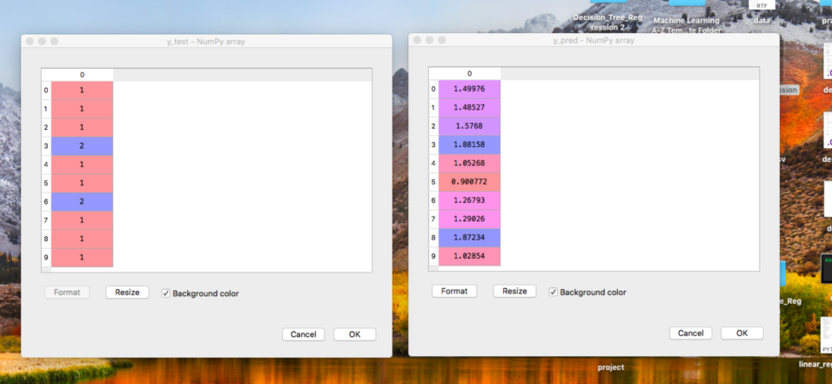


FIG.1

**ANALYSIS**: If the y\_pred value is greater than 1.5,it means that the person is having heart disease according to simple linear regression algorithm.

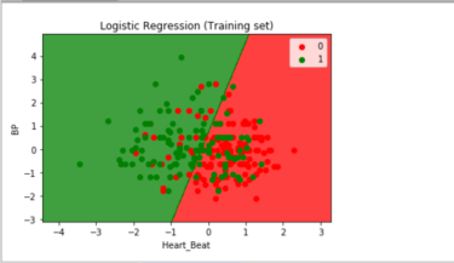


FIG.2

**Analysis:** The training set of is showed in graph that states if the person is having disease then he is lying above the fitted line and the one who is safe lies below the plotted line.



FIG.3

Analysis: This is the test data graph plot showing that the accuracy of 80 % of prediction.

**RESULT**

* ACCURACY OF LINEAR REGRESSION: 55.555
* ACCURACY OF MULTIPLE LINEAR REGRESSION:90
* ACCURACY OF LOGISTIC REGRESSION:80
* ACCURACY OF DECISION TREE:79.62
* ACCURACY OF SUPPORT VECTOR MACHINE:
* ACCURACY OF K-MEANS CLUSTERING:
* ACCURACY OF NAÏVE BAYES ALGORITHM:

**ANALYSIS**

We can say that among all these machine learning algorithms on the heart beat data set, multiple linear regression algorithm is showing the highest accuracy with maximum attributes .So the multiple linear regression is the best suitable among all the tested data for predicting the heart disease.

**FUTURE SCOPE**

Machine Learning has a huge potential and a whole fleet of applications. While some of the applications seemed quite far-fetched few years earlier, a lot has changed over the last couple of years. Engineers and scientists have attained major success in broadening the scope Artificial Intelligence and Machine Learning. Some leading tech giants like Google, Microsoft, IBM and Amazon are taking keen interest in Machine Learning and its applications. Amazon Inc. has recently launched its first Artificial Intelligence based Development Suite which also offers Machine Learning algorithms to the developers for leveraging maximum benefits out of the AI technology.

As an important part of healthcare is also the process a patient goes through and how they are treated. Improvements in this process can produce gains in both the quality and cost of care. I have had the opportunity to work on such projects with the Estonian Health Insurance Fund myself. For this end, a branch of data science called process mining has been utilized with quite a bit of success .Research has also been done to try and predict illness in the whole populations, but to my knowledge no such model is actively in use .The applications of AI will most likely gain more ground in the near future as at the beginning of 2017 100+ companies were providing such services. Unlike our commonwealth, health data in the US is locked in incompatible silo's, making the data very expensive to get at if possible at all. Even companies like Kaiser, with unified records, struggle to leverage ML on very unwieldy data infrastructure.

**CONCLUSION**

This paper explored and evaluated the heart disease prediction accuracy of various machine learning algorithms on heart disease data set. At the end we can say that the multiple linear regression is suitable and best recommended for the health sector and doctor bots which works on the principle of symptoms and past dataset.

**LEARNING OUTCOME**

This paper talks about the various machine learning techniques in python and implementation of these machine learning algorithms on health based dataset for which we have come to know various aspects of each algorithm like how the evaluation take place and how to solve the problems occurring while implementation. This paper also concludes that multiple linear regression is the best fit in these analyzed algorithms.

**REFERENCES**

Hailong Zhang, Wenyan Gan, Bo Jiang, "Machine Learning and Lexicon based Methods for Sentiment Classification: A Survey", IEEE, ISBN 978-1-4799-5727-9/14 \$31.00 © 2014.

Xing Fang, Justin Zhan, "Sentiment analysis using product review data", Fang and Zhan Journal of Big Data, vol. 2, no. 5, 2015.

Kaijie Guo, Liang Shi, Weilong Ye, Xiang Li, "A Survey of Internet Public Opinion Mining", IEEE, ISBN 978-1-4799-2030-3 114/\$31.00 ©2014.

Andrew, A. M. (2000). AN INTRODUCTION TO SUPPORT VECTOR MACHINES AND OTHER KERNEL-BASED LEARNING METHODS by Nello Christianini and John Shawe-Taylor, Cambridge University Press, Cambridge, 2000, xiii+ 189 pp., ISBN 0-521-78019-5 (Hbk,£ 27.50). Robotica, 18(06), 687-689.

Chen, M., Ebert, D., Hagen, H., Laramee, R. S., Van Liere, R., Ma, K. L., et al. (2009). Data, information, and knowledge in visualization. Computer Graphics and Applications, IEEE, 29(1), 12-19.