

# Predicting Heart Disease Using Machine Learning Tools and Techniques

C Ramya, Manoj Kumar D S

**Abstract**— Machine learning is closely associated with process statistics, that focuses on creating predictions by utilizing computers. The study of mathematical improvement delivers strategies, theory and application domains to the sector of machine learning. Data processing could be a field of study inside machine learning, and focuses on explorative information analysis through unsupervised learning. Demand for machine is growing tremendously to the explosion in data volume. IT industry is continuously inching forward to fulfil the demand by developing multiple tools and incorporating various techniques for machine learning process. This project deals with exploring and investigating different tools available in the IT market. Also, it deals with applying linear regression algorithm on heart disease dataset using various tools and techniques. As a result, we will be able to discuss the advantages and disadvantages of each practical tool that is being used in this project.

**Keywords:** Machine Learning, Supervised Learning, Logistic Regression Algorithm.

## I. INTRODUCTION

Machine learning is the logical investigation of calculations and factual models that PC frameworks use to continuously enhance their executions on the particular errand. Machine learning calculations fabricate a scientific model of test information, known as "preparing information", so to settle on forecasts or choices without being unequivocally modified to play out the undertaking. Machine learning calculations are utilized in the uses of email isolating, area of gate crashers, and PC vision, where it is infeasible to build up a calculation of explicit guidelines for playing out the undertaking. Machine learning is strongly identified with computational insights, which centres around making forecasts utilizing PCs. The investigation of numerical streamlining conveys techniques, speculation and application spaces to the field of machine learning. Information mining is a field of concentrate inside machine learning, and spotlights on exploratory data information examination through unsupervised learning. In its application crosswise over business issues, machine learning is additionally alluded to as prescient examination.

## II. OBJECTIVE

### Review Stage

The aim of this project is to find the capabilities of machine learning tools available in the market, to come up with the advantages and disadvantages of each tool based on an experimental analysis with data sources available from

the open source network. In this project, building predictive data model in the machine learning tool will be identified from major IT giants like Google, Microsoft and Oracle. Check the level of easiness offered to both the developers and as well to the end users (business users). The advantages and disadvantages of each of these tools will also be explored in terms of different features, different traditional ML algorithms and capability of integrating programming languages like R and Python. With multiple tools and technologies related to machine learning accessible in the market it is always a challenge to the developer or the architect team in IT industry to decide on the most efficient one. The best tool helps in building modules at high degree of accuracy in multiple domains (marketing, sales, HR, logistics, insurance, life sciences). On the other hand, the tools which are available in the market for machine learning will in turn help the business owner to use them with much ease without having a total dependency on the development team. The following are steps to execute this project.,

- Tools installation and configuration:
  1. DVD-Oracle
  2. Azure ML-Power BI
  3. Google Analytics
- Dataset exploration and explanation (EDA)
- ML model building.
- Tools performance evaluation.

## III. PROBLEM STATEMENT

Machine learning being a concept-oriented feature in the IT industry which serves people in predicting the data. The problem is that there are multiple programming languages, multiple tools involved in order to construct a prescient model. But the selection of the right tool or the right technology always helps the customers in building an efficient predictive model that gives more accurate outcome. Hence, choosing a right tool is a major task. Identifying the right tool for the right business is a problem. Complexity is one of the issues that is being encountered. As effortlessness may make an inadmissible reaction, it is likewise conceivable that, to speak to the multifaceted nature in information with an excessively intricate mapping. There can also be shortage of required resources. Due to shortage in resources the cost might be high.

**Revised Manuscript Received on April 05, 2019.**

**C Ramya**, UG Scholar, Department of CSE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, (Email: cramyasekar.97@gmail.com.)

**Manoj Kumar D S**, Assistant Professor, Department of CSE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, (Email: manojkumards03@gmail.com.)



#### IV. LITERATURE SURVERY

##### 4.1 Machine Learning- Applications

This investigation incorporates fundamental data about AI and recommender frameworks with their models. All the more extensively tended to was the point of AI's calculations, which are utilized in such frameworks. The paper primarily cantered around separating calculations relying upon the area of clients or questions, and dependent on substance. The portrayal of these calculations incorporates: likenesses, burdens and focal points, measures for the assessment of the calculation, and count of the example estimation of the assessment expectation. The plan some portion of the work starts with the portrayal of the utilized databases from the Movie Lens entrance. Thereafter, the innovation and the reasonable execution of the calculations depicted above are at that point exhibited. The following part contains an examination of the outcomes and ends dependent on the re-enactments completed on the PC to evaluate how the calculations work. At the completing phase of the work, there is a rundown, execution assessment of suggestion frameworks, and exercises gained from the venture, just as a proposition for further work on the issue of such frameworks. [2]

##### 4.2 Machine Learning- Diagnosing Breast Cancer

So as to the Breast Cancer Institute (BCI), Malignancy is a champion among the most perilous kind of infections that is successful for ladies on the planet. According to clinical master identifying this malignant growth in its first stage helps in sparing lives. It offers individualized assistants for more than 120 sorts of danger and related acquired clutters. For recognizing bosom malignancy generally AI methods are utilized. In this paper, it is recommended that versatile troupe casting a ballot technique for analyzed bosom disease utilizing Wisconsin Breast Cancer database. The point of this work is to think about and clarify how ANN and strategic calculation furnish better arrangement when its work with troupe AI calculations for diagnosing bosom malignancy even the factors are diminished. In this paper the utilization of Wisconsin Diagnosis Breast Cancer informational index is finished. At the point when contrasted with related work from the writing. It is demonstrated that the ANN approach with calculated calculation is accomplished 98.50% exactness from another AI calculation. [1]

##### 4.3 Multiple Instance- Learning of Social Network Depressed Users

A huge number of individuals are experiencing sorrow around the world. Dejection is considered as the most normally realized psychological well-being issue. By its inclination, sadness can reoccur. Numerous individuals experiencing despondency will in general lose intrigue, have low state of mind, feel sad, or have social disengagement. Even under the least favorable conditions, discouragement can prompt suicide. Up until now, there are a couple of quantities of studies researching significant learning systems to group interpersonal organization clients with dejection. A large portion of the investigations utilized established AI systems e.g., relapse, bolster vector machine, or choice trees. The target of this paper is to build up a profound learning

prescient model to order clients with misery. Since discouragement is an intermittent malady, it is fascinating in finding bizarre examples in client produced content after some time. Interpersonal organization posts after some time were removed for time arrangement. The prescient model for the arrangement was gotten from profound learning procedures. [4]

##### 4.4 Machine Learning- Detecting Brain Tumor

The medicinal pictures are the most critical assets utilized by specialists to analyze cerebrum tumors. An apparatus comprising with high exactness to mechanize the system can be staggeringly critical. In addition, because of the issues that are identified with legitimate liabilities, such a device can't be supplanted by the master assessments of prepared doctors. Here, a framework is intended to accurately group new cerebrum MRI pictures into pictures with tumor and pictures without tumor. This must be managed with no human mediation. To apply a few sorts of classifiers, we may likewise need to preprocess a few parts of the pictures, for example, the shading, zone of intrigue, picture file augmentation, and difference level. Later on, the most basic and segregating highlights of the preprocessed pictures were removed. In this stage, ten assortments of highlights are being separated. At long last, an instrument called WEKA 3.6 is utilized to apply four different classification calculations on these highlights and compute the accuracy/review, the F-measure, the level of effectively classified pictures and furthermore the length taken to manufacture each model. [6]

##### 4.5 Machine Learning- Fraudulent in Credit Cards

The purpose of data examination is to depict covered precedents and use them to help taught decisions in a collection of conditions. Charge card false is raising essentially including the headway of modernized innovation and turned into an obvious objective for cheats. Charge card misrepresentation has profoundly imbalanced open informational collections that are accessible. This paper, applies many regulated AI estimations to distinguish MasterCard deceitful exchanges utilizing a genuine informational index. Afterward, we utilize these calculations to actualize a super classifier utilizing group learning strategies. The most critical factors that may prompt higher precision in charge card counterfeit trade discovery is distinguished. Additionally, a correlation and exchange for the execution of various kinds of coordinated AI figuring that are existing in writing against the super classifier are actualized and did. [16]

## V. SYSTEM ARCHITECTURE

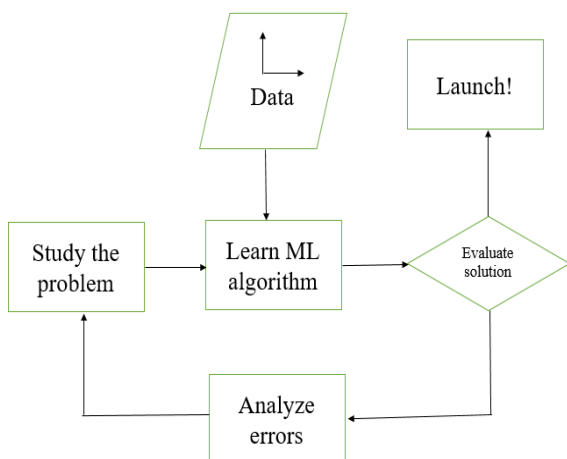


Fig. 1. Architecture Diagram

Fig No: 1 is the architecture of the system that will be carried out. Firstly, for predicting the heart disease dataset, it is important to study and understand the entire problem clearly. Then a proper machine learning algorithm has to be selected to carry on the implementation process. Then, the collected data undergoes the EDA [Exploratory Data Analysis] process to remove all the unwanted data. Then, this filtered data from the EDA is given to each tool by applying the machine learning algorithm and then the evaluation process is done. The linear regression algorithm is used for predicting the data. In the evaluation process the data is predicted if the patient has a coronary illness or not. If the data that is predicted is satisfactory then the final solution is launched else, it is resent to check for any errors. Finally, by determining the accuracy of each tool, there are possibilities to identify which tools provides better outputs.

### A. PROPOSED ALGORITHM

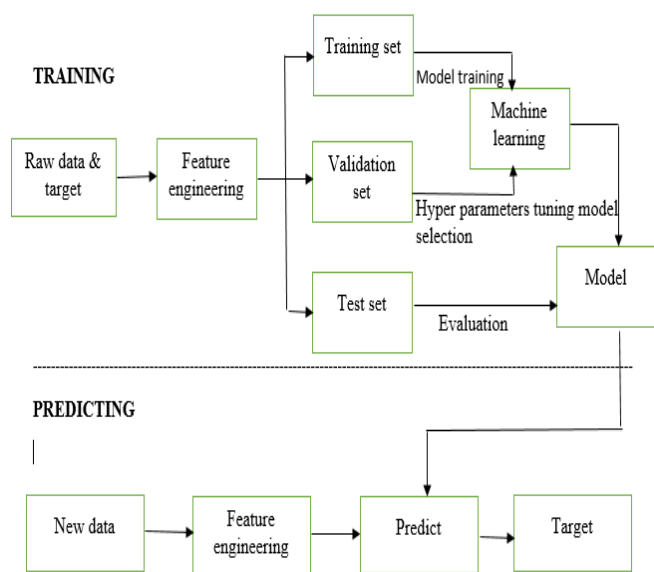


Fig. 2. Feature Engineering Process

The above block diagram (Fig No: 1.2) depicts that the raw data and the target is given to the process of feature engineering. Feature engineering is a process that mainly holds the domain knowledge and it brings out features in a

way that the machine learning works. If the process of feature engineering is done perfectly then it has the ability to increase the power of predicting the raw data in the machine learning. With the assistance of feature engineering predicting whether the model is successful or un-successful is done. The feature engineering data are classified as training set, validation set and test set. The training set and the validation set is given to the machine learning using model training and hyper parameter tuning model selection. The hyper parameter model tunes and configures the data in the machine learning accordingly. Once the data is tuned, the resultant data is sent to the model along with the testing data for the evaluation process and then the prediction the data is done.

### B. LINEAR REGRESSION ALGORITHM

Regression could be a methodology of modelling a target worth supported freelance predictors. This methodology is generally used for prognostication and looking for cause and impact relationship between variables. Regression techniques principally dissent supported the quantity of freelance variables and also the variety of relationship between the freelance and dependent variables. Simple regression may be a kind of multivariate analysis wherever the number of freelance variables is one and there's a linear relationship between the independent(x) and dependent(y) variable. The line within the on top of graph is stated because the best work line. supported the given knowledge points, we have a tendency to try and plot a line that models the points the most effective. the road is often modelled supported the equation shown below. The linear regression algorithm finds a best possible result for  $a_0$  and  $a_1$ .

$$\text{Equation for Linear Regression: } Y = a_0 + a_1 * x$$

## VI. RESULTS AND OUTPUT

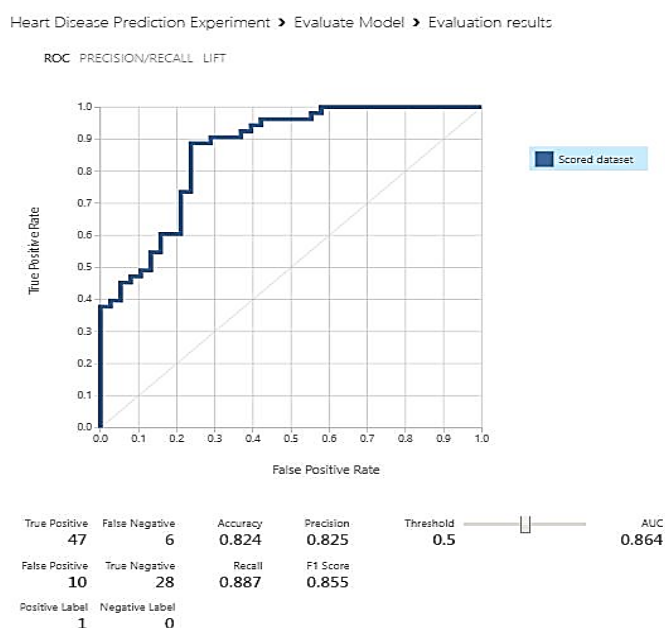


Fig. 3. Output from Azure ML Tool

The above figure (Fig 3.) is the output from Azure machine learning and Power BI tool. From this tool, the conclusion is that, it provides an accuracy of 82.4% for predicting the heart disease rate.

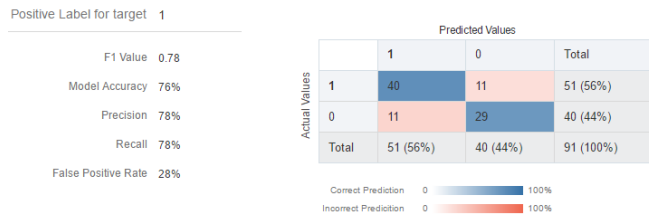


Fig. 4. Output from DVD-Oracle Tool

The above figure (Fig 4.) is the output from DVD-Oracle tool (Data Visualization Desktop). This tool provides an accuracy of 76% for predicting the heart disease rate.

By comparing both the tools it proves that Azure machine learning for Power BI tool provides a more accurate result. The hyper parameters tuning model selection is a superior choice for this system.

VII.CONCLUSION

These high-end tools really help and provides its customers with advanced features and concepts that drives machine learning projects with ease. Apart from providing concrete functions utilizing pre-built algorithms, these tools also offer a wide range of capabilities for exposing data in different visualizations and forms. Apart from the offerings of these tools in terms of utilizing the pre-built algorithms, it is necessary to explore about incorporating customized algorithms using python/R (or) any other statistical programming language. Doing so will help customers to utilize a same application for both customized algorithm intensive projects and as well as utilizing pre-built applications.

REFERENCES

1. B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICIC), Chennai, 2016.
2. G. Adomavicius, A. Tuzhilin „Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”.
3. Y. Tsehay et al., "Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI," 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, 2017, pp. 642-645.
4. S. L. Burcusa, W. G. Iacono, "Risk for recurrence in depression", *Clin. Psychol. Rev.*, vol. 27, no. 8, pp. 959-985, Dec. 2007.
5. G. Canbek and ü. Sa÷Öro÷lu, “A Review on Information, Information Security and Security Processes,” *Politek. Derg.*, vol. 9, no. 3, pp. 165– 174, 2006.
6. H. Najadat, Y. Jaffal, O. Darwish, and N. Yasser. A classifier to detect abnormality in CT brain images. In *The 2011 IAENG International Conference on Data Mining and Applications*, pages 374–377, Mar 2011.
7. R. M. Mohammad, F. Thabtah, and L. McCluskey, “Predicting phishing websites based on self-structuring neural

- network,” *Neural Comput. Appl.*, vol. 25, no. 2, pp. 443–458, 2014.
8. R. M. Mohammad, F. Thabtah, and L. McCluskey, “An assessment of features related to phishing websites using an automated technique,” *Internet Technol. ...*, pp. 492–497, 2012.
9. W. Hadi, F. Aburub, and S. Alhawari, “A new fast associative classification algorithm for detecting phishing websites,” *Appl. Soft Comput. J.*, vol. 48, pp. 729–734, 2016.
10. M. Ott, C. Cardie, and J. T. Hancock. Estimating the prevalence of deception in online review communities. In *ACM WWW*, 2012
11. K. Bennet & A.Demiris, “A Semi-supervised support Vector Machines” *Advances in Neural Information Processing Systems*,11, 368-374, 1999.
12. C. Campbell, N. Cristianini, and J.Shawe-Taylor, “Dynamically Adapting Kernels in Support Vector Machines”, *Advances in Neural Information Processing Systems*,Vol. 11. MIT Press, 1999 pp 204-210.
13. R Caruana, and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms” *Proceedings of the 23rd international conference on Machine learning*, 2006.
14. Corinna Cortes,Valdimir Vapnik, “Support-Vector Networks” *Machine Learning*, 20, 1995 pp 273-297.
15. G.Fung & O. Mangasarian, “Semi-supervised support vector machines for unlabeled data classification”. *Data Mining Institute, University of Wisconsin Madison (Technical Report 99-05)*
16. F. N. Ogwueleka, *Data Mining Application in Credit Card Fraud Detection System*, vol. 6, no. 3, pp. 311-322, 2011.
17. D.J.Miller and H.Uyar’s, “A mixture of experts classifier with learning based on both labelled and unlabelled data ” *Advance in NIPS* 9.571-577, 1997.
18. Oliver Chapelle, Bernhard Scholkopf, Alexandre Zien, “Semi supervised Learning [M],” *The MIT Press*, 2006