

Design & Implementation of Heart Disease Prediction using Machine Learning

Manoj S¹, Dr. Yuvaraju B.N²

¹Student, M.Tech-Information Technology, The National Institute of Engineering, Mysuru-09

²Professor, Dept. of CS&E, The National Institute of Engineering, Mysuru-09

Abstract—With the rampant increase in the heart strokes rates at younger ages, we need to put a system in place to be to detect the symptoms of a heart stroke at an early stage and thus prevent it. It is impractical for a common man to frequent undergo costly tests like the ECG and thus there needs to be system in place which is handy and at the same time reliable, in predicting the chances of heart disease. Thus we propose to develop an application which can predict the vulnerability of a heart diseases given basic symptoms like age, sex pulse rate, cholesterol etc. The machine learning algorithm neural network has proven to be most accurate and reliable algorithm and hence used in the proposed system.

Keywords— Machine Learning, neural network, Disease Prediction, Heart Disease Dataset

1. INTRODUCTION

As per the Centers for Medicare and Medicaid services, 50% of Americans have multiple chronic diseases with a total US health care expenditure in 2016 to be about \$3.3 trillion, which amounts to \$10,348 per person in the US. The early detection of common diseases such as breast cancer, diabetes, coronary artery and tumor, could control and reduce the chance of these diseases to be fatal for the patient. With the advancement in machine learning and artificial intelligence, several classifiers and clustering algorithms are being used to achieve this.

There is no dearth of records regarding medical symptoms of patients suffering heart strokes. However the potential they have to help us foretell similar possibilities in seemingly healthy adults are going unnoticed. For instance: As per the Indian Heart Association, 50% of heart strokes occur under 50 years of age and 25% of all heart strokes occur under 40 years of age in Indians.

Urban population is thrice as vulnerable to heart attacks as rural population. Thus propose to collect relevant data pertaining all elements related to our field of study, train the data as per the proposed algorithm of machine learning and predict how strong is there a possibility for a patient to contract a heart disease. Following the methodologies used in, this paper presents the use of machine learning algorithms for prediction of heart

diseases, which are the leading cause of deaths in the World.

The datasets used for the building the predictive models in this paper are available and can be downloaded from UCI machine learning library. The data is imported in CSV format and filtered for use. After data munging and attributes selection, machine learning algorithms including Logistic Regression, Decision Trees, Random Forest, Support Vector Machine(SVM) and K-Nearest Neighbour, Gradient Boosting, Naïve Bayes are used for prediction of the above-mentioned diseases, and a comparison of their accuracy is done for selecting best model for that disease dataset. All the analysis and visualization are carried out in python 3.1.

2. MACHINE LEARNING ALGORITHMS

2.1 Logistic Regression

Logistic Regression is a classification algorithm for the probability of occurrence of an event, whether that event will occur or not. It is used to portray a binary or a categorical outcome with only 2 classes. It is similar to linear regression with the only difference being that the outcome of the variable is categorical instead of a continuous variable. It uses Logit Link function, in which the data values are fitted, for prediction. The mathematical interpretation defines Logit function as the natural log of the odds that Y equals one of the categories. If p is the probability then, the logit function for p is defined as:

$$\text{Logit}(p) = \ln(p/1-p) \quad (1)$$

2.2 Decision Tree

Similar to the tree analogy in real life, the Decision tree is a machine learning algorithm, used for both classification and regression analysis. It is a tree-like graph beginning with a single node, and branching into its possible outcomes. Unlike the linear models, a decision tree is a supervised learning, that maps non-linear relationships as well. The data sample is divided into homogeneous subsets based on the most notable splitter in input attributes. The splitter is identified using various algorithms such as Gini Index, Chi-Square, Information Gain and Reduction in Variance. For example a dataset with boolean target variable, the entropy function for the dataset is given as:

$$\text{Entropy} = -p \log_2 p - p \log_2 p \quad (2)$$

2.3 Random Forest

Random forest is an ensemble of various decision trees, trained with the bagging methodology. Bagging is used for making the model more stable and accurate by approaching averaging model technique. The random forest classifier is basically a collection of decision tree classifiers where each tree is constructed with a number of random vectors and is able to vote for the most favored class for prediction. The injection of randomness in the model prevents it from over fitting and provide better result for classification analysis.

2.4 SupportVector Machine

SupportVector Machines, also called Support Vector networks are supervised learning algorithms used for both classification and regression analysis. It classifies the data points plotted in a multidimensional space into categories by parallel lines called the hyperplane. The classification of data points involves the maximization of margin between the hyperplane. There are different kernels available for mapping of linear or no linear data points in a multidimensional space for separation. For our analysis, we have used only the Linear and Radial basis function as kernel.

2.5. Navie Bayes (NB)

The Navie Bayes classifier is based on Bayes Theorem. In this classifier the independency between the attributes of the data set is the main assumption and the key point in order to make a prediction. It is easy to implement and particularly useful for very large data sets. In addition to its simplicity, this model is shown to outperform even highly sophisticated classification methods.

2.6 K-Nearest Neighbour

K-Nearest Neighbors classifies an object by the majority vote of its closest neighbors. In other words, based on some distance metrics, the class of a new instance will be predicted. The distance metric used in nearest neighbor methods for numerical attributes can be a simple Euclidean distance.

2.7 Gradient Boosting

A Gradient Boosting Classifier is used for classifying and predicting the heart and stroke disease from the extracted attributes. Experimental evaluation is carried out using Statlog heart disease dataset and International Stroke Trial Database on the factors such as classification accuracy, classification time, error rate and true positive rate with respect to number of patients.

3. PROPOSED METHOD

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could see it as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering.

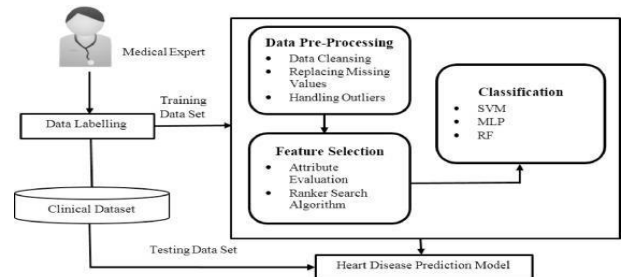


Fig-1: System Architecture

3.1 Data Access Layer

Data access layer is the one which exposes all the possible operations on the data base to the outside world. It will contain the DAO classes, DAO interfaces, POJOs, and Utils as the internal components. All the other modules of this project will be communicating with the DAO layer for their data access needs.

3.2 Account operation

- Account Operations
- Account operations module provides the following functionalities to the end users of our project.
- Register a new seller/ buyer account
- Login to an existing account
- Logout from the session
- Edit the existing Profile
- Change Password for security issues
- Forgot Password and receive the current password over an email
- Delete an existing Account
- Account operations module will be reusing the DAO layer to provide the above functionalities.

3.3 Implementation of sequential Model Algorithm

The sequential model is a theory that describes cooperativity of protein subunits. It postulates that a protein's conformation changes with each binding of a ligand, thus sequentially changing its affinity for the ligand at neighboring binding sites. This model for Allosteric regulation of enzymes suggests that the subunits of multimeric proteins have two conformational states. The

binding of the ligand causes conformational change in the other subunits of the multimeric protein. Although the subunits go through conformational changes independently (as opposed to in the MWC model), the switch of one subunit makes the other subunits more likely to change, by reducing the energy needed for subsequent subunits to undergo the same conformational change.

3.4 Training and Testing the model for accuracy

Here, the model will be trained using the datasets and tested for finding the accuracy of the model. Optimization will be done to improve the accuracy if needed. In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data. Such algorithms work by making data-driven predictions or decisions, through building a mathematical model from input data. The data used to build the final model usually comes from multiple datasets. In particular, three data sets are commonly used in different stages of the creation of the model.

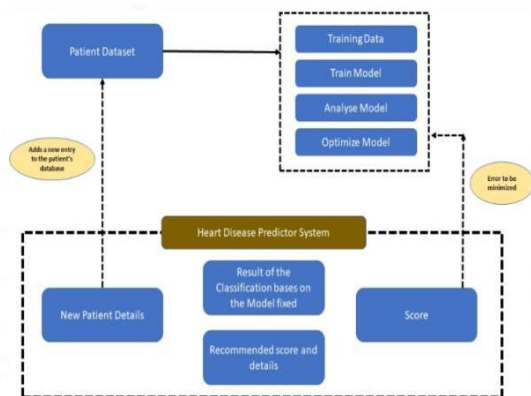


Fig-2: Data Flow

The above fig 2 shows the model is initially fit on a training dataset, that is a set of examples used to fit the parameters (e.g. weights of connections between neurons in artificial neural networks) of the model. The model (e.g. a neural net or a naïve Bayes classifier) is trained on the training dataset using a supervised learning method (e.g. gradient descent or stochastic gradient descent). In practice, the training dataset often consist of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), which is commonly denoted as the target (or label). The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation. Successively, the fitted model is used to predict the responses for the observations in a second dataset

called the validation dataset. The validation dataset provides an unbiased evaluation of a model fit on the training dataset while tuning the model's hyper parameters (e.g. the number of hidden units in a neural network). Validation datasets can be used for regularization by early stopping: stop training when the error on the validation dataset increases, as this is a sign of over fitting to the training dataset. This simple procedure is complicated in practice by the fact that the validation dataset's error may fluctuate during training, producing multiple local minima. This complication has led to the creation of many ad-hoc rules for deciding when over fitting has truly begun. Finally, the test dataset is a dataset used to provide an unbiased evaluation of a final model fit on the training dataset. If the data in the test dataset has never been used in training (for example in cross-validation), the test dataset is also called a holdout dataset.

3.5 Implementation of RESTful APIs for exposing the model to other apps/clients

Here, the APIs will be developed so that the existing applications can re-use the model we developed in the second module. Representational state transfer (REST) is a software architectural style that defines a set of constraints to be used for creating Web services. Web services that conform to the REST architectural style, called RESTful Web services, provide interoperability between computer systems on the Internet. RESTful Web services allow the requesting systems to access and manipulate textual representations of Web resources by using a uniform and predefined set of stateless operations. Other kinds of Web services, such as SOAP Web services, expose their own arbitrary sets of operations.

3.5 Cloud based deployment process of the model

Here, the model will be deployed on a cloud server to make the solution accessible across the geographical areas. For the cloud deployment process, I use Amazon web service.

Here is the overall representation of the project

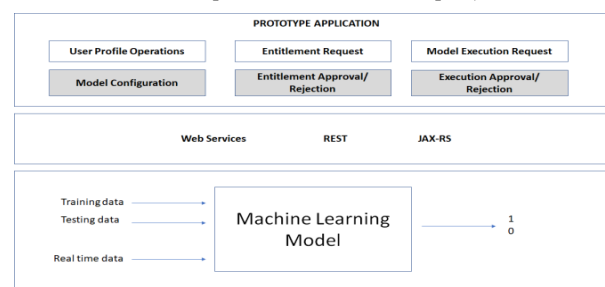


Fig-3: Model Representation

3.6 REST implementation

This module provides an end point to the outside world so that the third party applications can invoke it by sending the new patients' data. The end point upon receiving the data, will then invoke the seven python programs by sending the data as a command line parameters. Each of these python program will either give output as 1 or 0. The output 1 indicates there's a disease and the output 0 indicates there's no disease. The end point will then analyses the output from each of the models. The end point will count the number of models providing the output 0 and then the number of models providing the output 1. Whichever is having a majority, the model will consider that result as the actual result and then responds back the same to the third party application who invoked this. This REST web service endpoint is exposed as a POST method.

4. Experiment with Heart Disease Dataset

The Heart Disease Dataset consists of 14 input attributes including age, gender, type of chest pain, blood pressure, cholesterol, blood sugar level, electrocardiograph result, maximum heart rate, exercise-induced angina, old peak, Slope, number of vessels colored, thal. The dataset contains 303 instances, from which 2 instances for a number of vessels colored attribute and 4 instances for thal attribute are missing, which are filled by their mean value for the dataset respectively. The prediction attribute consists of 2 classes ranging from integer value 0 - 1 where 0 indicate no heart disease and the integer value from 1 indicate the presence of heart disease. The feature selection from 14 input parameter by backward elimination resulted in a total of 14 significant input parameters which include gender, type of chest pain, blood pressure, blood sugar level, electrocardiograph result, maximum heart rate, exercise induced angina, old peak, Slope, number of vessels colored, thal . Fig. 4 shows the Variation of age for each target class of the models obtained for heart dataset. Logistic Regression was found to have the highest accuracy among all.

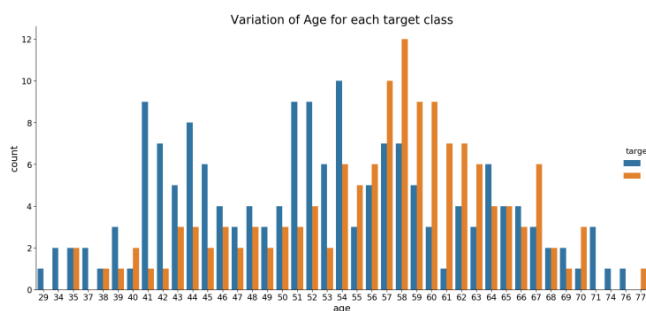


Fig-4: Variation of age for each target class

On the same metric, then compared the accuracy of our model with the SVM based model , and Pruning Decision

Tree method and found it to perform better. Table 1, present all the classification accuracies achieved by the algorithms following our proposed model.

TABLE-1: RESULTS

Classifier	Training Set Results	Test Set Results
Logistic Regression	86.36	80.32
Decision Tree	100	77.04
Random Forest	98.76	75.40
Support Vector Machine(SVM)	92.56	80.32
Gradient Boost	83.87	72.66
Navie Bayes (NB)	86.77	80.32
K-Nearest Neighbour	85.36	80.89

5. SCREENSHOTS

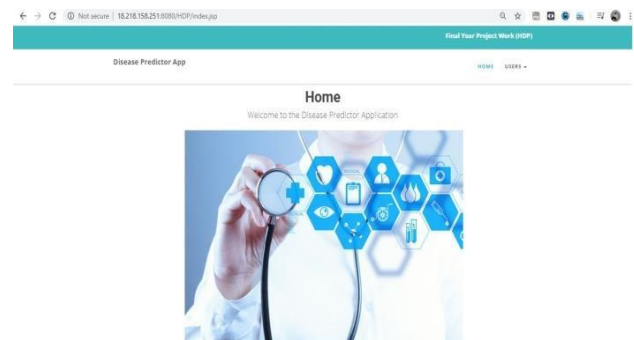


Fig-5: Home Page of Heart Disease Predictor Application

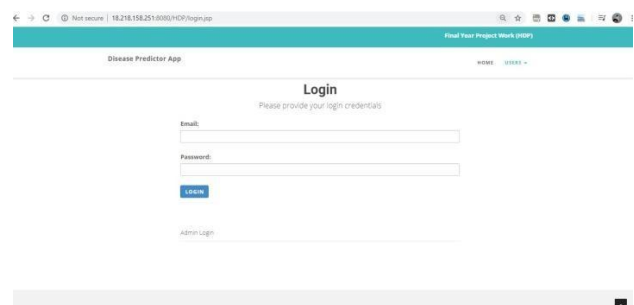
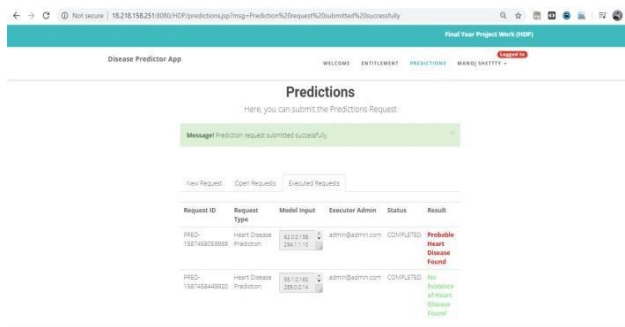


Fig-6: User Login Page



Request ID	Request Type	Model Input	Executor Admin	Status	Result
PRED-1587468239992	Heart Disease Prediction	43.0218 28.1712	admin@bairm.com	COMPLETED	Probable Heart Disease Found
PRED-158746848922	Heart Disease Prediction	43.0218 28.1712	admin@bairm.com	COMPLETED	Probable Heart Disease Found

Fig-7: Prediction Executed Results

6. CONCLUSION

The Heart Disease Prediction System using numerous Machine learning algorithm, viz. with a prediction result that gives the state of a user leading to diagnostics. Due to the recent advancements in technology, the machine learning algorithms are evolved a lot and hence we use multiple algorithms in the proposed system because of its efficiency and accuracy. Also, the algorithm gives the nearby reliable output based on the input provided by the users. If the number of people using the system increases, then the awareness about their current heart status will be known and the rate of people dying due to heart diseases will reduce eventually.

In Future, we aim to work on other disease prediction algorithms like cancer detection, retinopathic diabetes prediction, etc so that all the health related diagnostic can be obtained under a single platform.

REFERENCES

[1] "UCI Machine Learning Repository." [Online]. Available: <https://archive.ics.uci.edu/ml/index.php>. [Accessed: 21-Apr-2018].

[2] W. Bergerud, "Introduction to logistic regression models with worked forestry examples: biometrics information handbook no. 7," no. 7, p. 147, 1996.

[3] S. Sperandei, "Lessons in biostatistics Understanding logistic regression analysis," *Biochem. Medica*, vol. 24, no. 1, pp. 12–18, 2014.

[4] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[5] T. M. Mitchell, "Decision Tree Learning," *Machine Learning*, pp. 52–80, 1997.

[6] L. Breiman, "Random Forest," pp. 1–33, 2001.

[7] M. Denil, D. Matheson, and N. De Freitas, "Narrowing the Gap: Random Forests In TheDenil, M., Matheson, D., & De Freitas, N. (2014). Narrowing the Gap: Random Forests In Theory and In Practice. Proceedings of The 31st International Conference on Machine Learning, (1998), 665–673.Retrieved from ht," *Proc. 31st Int. Conf. Mach. Learn.*, no.1998, pp. 665–673, 2014.

[8] V. Jakkula, "Tutorial on Support Vector Machine (SVM)," *Sch. EECS, Washingt. State Univ.*, pp. 1–13, 2006.

[9] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," vol. 22, no. 2, pp. 103–104, 2000.

[10] Y. Freund and R. Schapire, "A Tutorial on Boosting," pp. 1–35, 2013.

[11] R. Rojas, "AdaBoost and the Super Bowl of Classifiers A Tutorial Introduction to Adaptive Boosting," *Writing*, pp. 1–6, 2009.

[12] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." [Online]. Available: [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisc+Onsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisc+Onsin+(original)). [Accessed: 21-Apr-2018].

[13] C. Sirisomboonrat and K. Sinapiromsaran, "Breast Cancer Diagnosis Using Multi-Attributed Lens Recursive Partitioning Algorithm," *2012 Tenth Int. Conf. ICT Knowl. Eng.*, pp. 40–45, 2012.

[14] D. Lavanya and D. K. U. Rani, "Analysis of Feature Selection with Classification : Breast Cancer Datasets," *Indian J. Comput. Sci.Eng.*, vol. 2, no. 5, pp. 756–763, 2011..

[15] "UCI Machine Learning Repository: Pima Indians Diabetes." [Online]. Available: <https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/>. [Accessed: 21-Apr-2018].

[16] R. Priyadarshini, N. Dash, and R. Mishra, "A Novel approach to Predict Diabetes Mellitus using Modified Extreme Learning Machine," *Int. Conf. Electron. Commun. Syst. (ICECS), IEEE*, pp. 1–5, 2014.