Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4571 - 4577
Received 25 April 2021; Accepted 08 May 2021.

GUI based Prediction of Heart Stroke Stages by finding the accuracy using Machine Learning algorithm

Yash Prakash Kadtan[1], Aditya Pratap Singh Chauhan[2], R. Brindha[3]
[1,2,3]SRM Institute of Science and Technology, Kattankulathur
yp9782@srmist.edu.in

ABSTRACT

Many predictive techniques are used and applied in the medical domain such as predicting occurrence, evaluating outcome of diseases and assisting clinicians to recommend treatment of diseases. Standard predictive models or methods, on the other hand, are incapable of simulating the complexities of feature representation in medical problem domains, and therefore are ineffective in capturing the underlying information. To address this problem, machine learning algorithms are used to apply predictive computational techniques for heart stroke on a given hospital dataset. Atrial fibrillation is a significant risk factor for cardiac attack in patients, and it shares many of the same factors that predict stroke. When a dataset is analysed using a controlled machine learning algorithm, variables such as variable recognition, univariate analysis, bivariate and multivariate analysis, missed value therapies, mathematical methods, and so on are all recorded. The aim of the predictive analytics model is to recognise the various stages of heart stroke in patients. Discuss the output of the provided hospital dataset, as well as the evaluation of the classification study and the uncertainty matrix. To compare supervised classification machine learning algorithms and suggest a machine learning-based method for reliably predicting heart stroke using given characteristics. Furthermore, compare and discuss the performance of different machine learning algorithms from the given healthcare department dataset with evaluation classification report, define the confusion matrix, and categorise data from priority, and the result depicts that the effectiveness of a graphical user interface based proposed machine learning algorithm technique can be compared with best accuracy with precision and F1 Score

Keywords
Dataset, python, Prediction of Accuracy result.

Introduction

The goal of machine learning is to forecast long-term data based on historical data. ML is a branch of AI that allows computers to understand and learn without having to be specifically programmed. Machine learning basics, which involve the introduction of a basic machine learning algorithm in Python, focuses on the development of computer programmes that can respond to new data. Specialized algorithms are used in the coaching and prediction process. It feeds the training data to an algorithm, which then makes forecasts on the substituted test data using the training data. Machine learning is often divided into three categories. Supervised, unsupervised, and reinforcement learning are the three forms of learning. Both the input file and the accompanying labelling to locate data in supervised learning must be labelled by a person's being beforehand. There is no mark for unsupervised learning. It was in charge of the algorithm for learning. The input file's clustering must be determined by this algorithm. Finally, Reinforcement learning interacts with its environment in a dynamic way, providing positive or negative feedback to help it improve.

Using data processing and machine learning approaches to predict disease based on patient treatment history and health data has been a struggle for decades. Many studies have used data mining methods to predict particular diseases using pathological data or medical profiles. These methods are aimed at predicting disease recurrence. In addition, other methods seek to predict disease control and progression. Machine learning's recent performance in a number of fields has sparked a move toward models that can learn rich, hierarchical representations of raw data with

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4571 - 4577
Received 25 April 2021; Accepted 08 May 2021.

little pre-processing and produce more accurate results. Several papers on data mining strategies for diagnosis of heart disease, such as Decision Tree, Naive Bayes, and support vector machines, have been written, with varying degrees of accuracy in disease prediction.

Literature Review

The effectiveness of treating therapies for cardiac associated shock and heart failure is regulated or limited by limitations in available diagnostic metrics. Measurement of pulmonary capillary wedge pressure and dependence on linear approximations between pressure and flow to measure peripheral vascular resistance are used in existing clinical procedures to infer cardiovascular state. What makes this method more effective is the use of the right set of algorithms, which can begin with data selection, cleaning, and ending with the development of a classification model. The effectiveness of various algorithms such as Logistic Regression, Naive Bayes, kNN, random forest, SVM, and decision tree is the subject of this article.

Title	Objective	Disadvantages
A Machine Learning Approach to Classifying Self-Reported Health Status in a cohort of Patients with Heart Disease using Activity Tracker Data	A temporal machine learning model was used to classify self-reported physical health in patients with SIHD using physiological indices measured by activity trackers.	Less accurate because of insufficient data
Heart Disease Prediction using Evolutionary Rule Learning	Heart disease prediction system was proposed to identify the risk of heart disease accurately. To generate strong association rules, the model has shown frequent pattern growth association mining on patient's dataset.	Not a reliable Model
Sparse Support Vector Machine for Intrapartum Fetal Heart Rate Classification	This contribution aims to promote Sparse Support Vector Machine classification that allows both to select a small number of relevant features within a large list and to achieve efficient fetal acidosis detection.	The model mainly relies on the pH level(measured from post-birth umbilical cord artery blood sample). The dataset was very small.

Topic

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4571 - 4577
Received 25 April 2021; Accepted 08 May 2021.

Machine learning (ML) is an artificial intelligence (AI) technique that allows computers to learn without being explicitly programmed.Machine learning is concerned with the creation of computer programs that can adapt to new data, as well as the fundamentals of machine learning, such as the construction of a simple machine learning algorithm in Python.Specialized algorithms are used in the training and prediction process.A supervised learning algorithm is provided both the input data and the accompanying labelling to learn data, which must first be tagged by a person.Finally, reinforcement learning interacts with its surroundings in a dynamic manner and gets positive or negative feedback in order to enhance its performance.

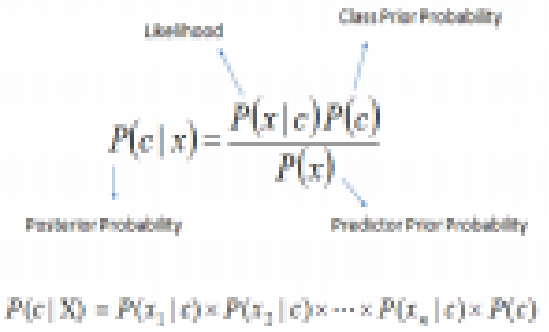
Methods

A. Logistic Regression

It's a mathematical approach for analysing a collection of data in which one or more independent variables affect the outcome. A dichotomous attribute is used to test the results (there are only two possible outcomes). The aim of logistic regression is to find the model that best represents the relationship between a series of independent variables and an interesting dichotomous function (dependent variable equals outcome variable). Logistic regression can be used to quantify the likelihood of a categorical variable. The vector in logistic regression may be a binary variable with data coded as 1 (success) or 0 (failure) (failure).

B. Naïve Bayes

Naive Bayes algorithm based on Bayes' theorem assumes independence and equality among its features. This classifier is very useful in real life applications such as document classification.



This algorithm is able to estimate parameters on the basis of small training data. These classifiers are known to be faster compared to more sophisticated methods.

C. K Nearest Neighbour

kNN algorithm can be used for both classification and regression. Under classification, the output is a labelled class. A particular data point is classified on the basis of majority votes of its nearest k neighbours.

This algorithm is particularly more advantageous when it comes to more noisy and larger data sets as it is able to handle them easily.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4571 - 4577
Received 25 April 2021; Accepted 08 May 2021.

D. Random Forest

Random Forest uses central tendencies man and mode to classify a dataset. This algorithm can be used for classification as well as regression. Multiple decision trees are created in this algorithm at the time of training and for classification, the mode is the output and for regression the mean of the trees is the output. This algorithm usually results in the highest accuracy and is also able to handle big data efficiently. It often balances datasets automatically and has methods for missing data problems.

E. Support Vector Machine

Support Vector Machine results in a hyperplane clearly classifying the given data points into different groups of similar features. The new data points are then assigned to one of the already existing groups on the basis of similarity. This algorithm is easily able to handle high dimensional data with more memory efficiency as it uses the training data as a subset for decision making.

F. Decision Tree

A decision tree is a classification or regression model in the form of a tree structure. It separates the data set into smaller subsets while building a decision tree at the same time[7]. A classification or preference is represented by a leaf node, and a decision node has at least two or more branches. The topmost node in a tree that corresponds to the most effective predictor is called a root node. Decision trees can be used to control both categorical and numerical data. A decision tree is a type of tree structure used in regression modelling[8].

Methodology

The data used for this paper was taken from a Kaggle repository and it is imbalanced. The data has various labelled features such as gender, age, if a person has hypertension or not, if a person has heart disease or not, if a person is married or not, work type, residential type, average glucose level, BMI, smoking status, if a person has a stroke or not. There are a total of twelve features in which four are numeric, four alphanumeric and the rest categorical. The final column is the target column with targets 1 and 0.

Data Analysis

Data visualisation is one of the most useful capabilities in applied statistics and machine learning. Statistics is mainly concerned with the description and calculation of quantitative results. Furthermore, data visualisation offers a useful range of methods for gaining a qualitative understanding. When discussing and learning about a dataset, as well as finding patterns, corrupt results, and outliers, this can be useful. In addition to indicators of affiliation and meaning, data visualisations should be used for domain awareness to communicate and clarify critical interactions in more visceral and stakeholder-friendly plots and maps.Sometimes data does not make sense until it can be looked at in a visual form, such as with charts and plots. Both applied analytics and applied machine learning include the ability

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4571 - 4577
Received 25 April 2021; Accepted 08 May 2021.

to quickly visualise data and other samples. It will go through different plot styles that you'll need to know while visualising data in Python, as well as how to use them to help interpret your own data.

➢How to use categorical quantities and line plots with bar charts to map time series results.

➢How to use box plots and histograms to summarize data distributions.

➢How to use scatter plots to create a relationship between variables.

B. Data Pre-processing

The transformations we apply to our data before feeding it to the algorithm are referred to as pre-processing. The procedure of transforming raw data into clean data sets is known as information preprocessing. To put it another way, once data is obtained from various sources, it is doing so in a raw format, making interpretation difficult. To get better outcomes from the implemented model in the Machine Learning methodology, the data must be in the right format. Some Machine Learning algorithms need data in a particular format; for example, the Random Forest algorithm does not allow null values. As a consequence, in order to run random forest algorithms, null values from the original raw data set must be managed. Another factor to note is that data sets can be organised in such a way that they can be used by different Machine Learning algorithms.

C. Accuracy Calculation

There are four types of results that can arise while making classification predictions.

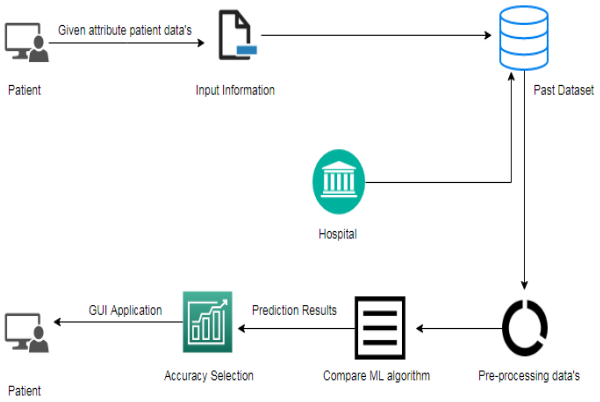
- True positives occur when you predict that an observation belongs to a group and it really does.
- True negatives: when you predict that an observation does not belong in a category, it does not belong in that category.
- False positives happen when you think an observation belongs to a group when it doesn't.
- False negatives happen when you incorrectly predict that an observation does not belong in a group when it really does.

The three main metrics for a classification model are accuracy, precision, and recall.

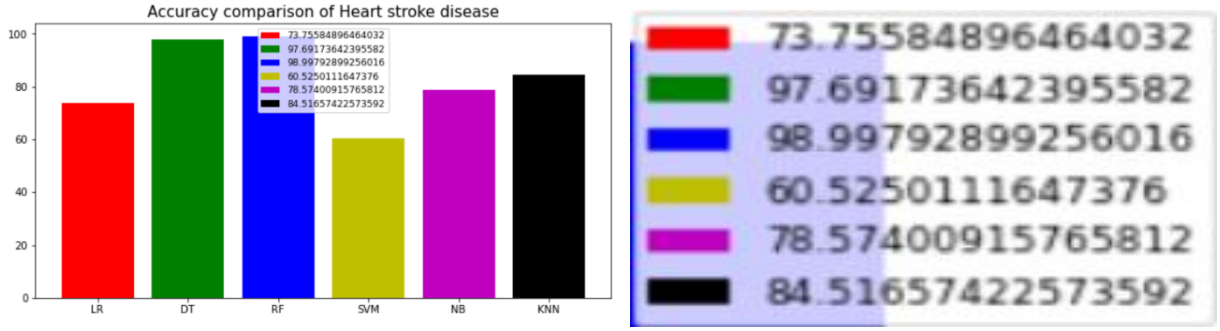
- Accuracy - The percentage of accurate test data forecast. It's usually easy to calculate by dividing the number of correct predictions by the total number of predictions.
- Precision - The fraction of relevant examples (true positives) among all of the examples which were predicted to belong during a certain class.
- Recall - The fraction of examples which were predicted to belong to a category with reference to all of the examples that really belong within the class.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4571 - 4577
Received 25 April 2021; Accepted 08 May 2021.

D. Architecture diagram



Results



Discussions

Decision Tree and Random Forest both are very reliable algorithms and both of them have performed very well in our model. Random forest is basically a combination of many decision trees and thus most of the times it outperforms Decision Tree by a somewhat better accuracy.

Conclusion

As a result, DT and RF are strong algorithms that provide the highest accuracy in our application. But Random forest turns out to be the best. The patient's best chance of minimizing the effects of astroke is to have a diagnosis as soon aspossible.To present an artificial intelligence-assisted predictionmodel that outperforms human accuracy while still allowing for early detection.

Limitations and Future Studies

Hospital wants to automate the detecting of the heart stroke from eligibility process (real time) based on the account detail.To automate this process by show the prediction result in web

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4571 - 4577
Received 25 April 2021; Accepted 08 May 2021.

application or desktop application. To optimize the work to implement in Artificial Intelligence environment.

Acknowledgement

I would like to thank my guide R. Brindha as well as my research partner Aditya Pratap Singh who helped me a lot during the research and I got to learn so many new things which wouldn't have been possible without them.

References

[1] Yiwen Meng, William Speier, Member, Chrisandra Shufelt, Sandy Joung, Jennifer E Van Eyk, C. Noel Bairey Merz, Mayra Lopez, Brennan Spiegel and Corey W. Arnold, 2019, "A Machine Learning Approach to Classifying Self Reported Health Status in a cohort of Patients with Heart Disease using Activity Tracker Data"

[2] Evanthia E. Tripoliti, Member, Penelope Ioannidou, Petros Toumpaniaris, Aidonis Rammos, Dominique Pacitto, Jean-Christophe Lourme, Yorgos Goletsis, Member, Katerina K. Naka, Abdelhamid Errachid, Dimitrios I. Fotiadis, Fello, 2019, "Point-of-care testing devices for heart failure analyzing blood and saliva samples"

[3] Lin Zhang, Senior Member, Heng Chen, Senior Member, Qiushi Wang, Member, Neeraj Nayak, Member, Yanfeng Gong, Senior Member, and Anjan Bose, Fellow, 2019, "A Novel On-line Substation Instrument Transformer Health Monitoring System Using Synchrophasor Data"

[4] Qiu Xiao, Jiawei Luo, and Jianhua Dai , 2019, "Computational Prediction of Human Disease Associated circRNAs based on Manifold Regularization Learning Framework"

[5] Paulo H. Oliveira, Lucas C. Scabora, Mirela T. Cazzolato, Willian D. Oliveira, Rafael S. Paixao, Agma J. M. Traina, and Caetano Traina-Jr, 2018, "Employing Domain Indexes to Efficiently Query Medical Data from Multiple Repositories"

[6] Ganapati Bhat, Ranadeep Deb, Umit Y. Ogras, 2019, "OpenHealth: Open Source Platform for Wearable Health Monitoring"

[7] Utkarsh Konge, Abhishek Baikadi, Jayanth Mondi, Sivakumar Subramanian. "Data-Driven Model Based Computation and Analysis of Operability Sets Using High-Dimensional Continuation: A Plant-Wide Case Study" , Industrial & Engineering Chemistry Research, 2020

[8] Pal, Priya, and Deepak Motwani. "The Based on Rough Set Theory Development of Decision Tree after Redundant Dimensional Reduction" 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015.