NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM,

APPROVED BY AICTE & GOVT.OF KARNATAKA

LA PROPOSAL

Comparison of Various Machine Learning Algorithms on UCI Heart Disease Dataset

Submitted in partial fulfilment of the requirement for the LA Component

Introduction to Machine Learning Course (18CSE71) *of 7th Semester*

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

Karan R	1NT18CS067
Samiksha Ullal	1NT18CS143

Course Coordinator
Dr. Vani V
Professor, Dept. of CS&E, NMIT

Department of Computer Science and Engineering

ABSTRACT

Heart disease is one of the most significant causes of mortality in today's world. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis, with the current epidemic scenario doctors need a support system for more accurate prediction of heart disease. Machine learning algorithm opens new door opportunities for precise prediction of heart diseases. Diagnosis and prediction of heart related diseases requires more precision, perfection and correctness because a little mistake can cause havoc or death of the person. To deal with this problem there is an essential need of prediction system for awareness about diseases. Machine learning is the branch of Artificial Intelligence, it provides prestigious support in predicting any kind of event which take training from natural events. Our objective is to calculate accuracy of various machine learning algorithms for predicting heart disease, algorithms used are k-nearest neighbor, decision tree, logistic regression, random forest, Naïve bayers and support vector machine (SVM). We are the existing dataset from the Cleveland database of UCI repository of heart disease patients.

INTRODUCTION

Machine learning is one of the most rapidly evolving domains of artificial intelligence. Machine learning algorithms can analyze huge data from various fields, one such important field is the medical field. Its primary focus is to design systems, allow them to learn and make predictions based on the experience. It trains machine learning algorithms using a training dataset to create a model. The model uses the new input data to predict heart disease. Using machine learning, it detects hidden patterns in the input dataset to build models. It makes accurate predictions for new datasets. The dataset is cleaned and missing values are filled. The model uses the new input data to predict heart disease and then tested for accuracy.

The machine learning methods that will be used to predict the heart disease are Logistic regression, Naïve bayers, Support vector machine (SVM) ,K-NN, decision tree and random forest.

DATASET

Characteristics of the dataset:

Data Set Characteristics	**Multivariate**
Attribute Characteristics	Categorical, Integer, Real
Associated Tasks	Classification
Number of Instances	303
Number of Attributes	75
Missing Values	Yes
Area	Life

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. 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. Experiments with the Cleveland database have concentrated on simply

attempting to distinguish presence (values 1,2,3,4) from absence (value 0).
The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.
One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.
To see Test Costs (donated by Peter Turney), please see the folder "Costs".

Attribute Information:
Only 14 attributes used:

S.No	Attribute	Description	Type
1.	Age	Patient's age (29 to 77)	Numerical
2.	Sex	Gender of patient(male-0 female-1)	Nominal
3.	Cp	Chest pain type	Nominal
4.	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5.	Chol	Serum cholesterol (in mg/dl, values from 126 to 564)	Numerical
6.	Fbs	Fasting blood (sugar>120 mg/dl, true-1 false-0)	Nominal
7.	Resting	Resting electrocardiographic result (0 to 1)	Nominal
8.	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9.	Exang	Exercise (1-yes 0-no)	Nominal
10.	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11.	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12.	Ca	Number of major vessels (0-3)	Numerical
13.	Thal	3-normal	Nominal
14.	Targets	1 or 0	Nominal

Complete attribute documentation:
1. id: patient identification number
2. ccf: social security number (I replaced this with a dummy value of 0)
3. age: age in years
4. sex: sex (1 = male; 0 = female)
5. painloc: chest pain location (1 = substernal; 0 = otherwise)
6. painexer (1 = provoked by exertion; 0 = otherwise)
7. relrest (1 = relieved after rest; 0 = otherwise)
8. pncaden (sum of 5, 6, and 7)

9. cp: chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic

10. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
11. htn
12. chol: serum cholestoral in mg/dl
13. smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)
14. cigs (cigarettes per day)
15. years (number of years as a smoker)
16. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
17. dm (1 = history of diabetes; 0 = no such history)
18. famhist: family history of coronary artery disease (1 = yes; 0 = no)
19. restecg: resting electrocardiographic results

    -- Value 0: normal
    -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

20. ekgmo (month of exercise ECG reading)
21. ekgday(day of exercise ECG reading)
22. ekgyr (year of exercise ECG reading)
23. dig (digitalis used furing exercise ECG: 1 = yes; 0 = no)
24. prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
25. nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)
26. pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
27. diuretic (diuretic used used during exercise ECG: 1 = yes; 0 = no)
28. proto: exercise protocol

    1 = Bruce
    2 = Kottus
    3 = McHenry
    4 = fast Balke
    5 = Balke
    6 = Noughton
    7 = bike 150 kpa min/min (Not sure if "kpa min/min" is what was written!)
    8 = bike 125 kpa min/min
    9 = bike 100 kpa min/min
    10 = bike 75 kpa min/min
    11 = bike 50 kpa min/min
    12 = arm ergometer

29. thaldur: duration of exercise test in minutes

30. thaltime: time when ST measure depression was noted
31. met: mets achieved
32. thalach: maximum heart rate achieved
33. thalrest: resting heart rate
34. tpeakbps: peak exercise blood pressure (first of 2 parts)
35. tpeakbpd: peak exercise blood pressure (second of 2 parts)
36. dummy
37. trestbpd: resting blood pressure
38. exang: exercise induced angina (1 = yes; 0 = no)
39. xhypo: (1 = yes; 0 = no)
40. oldpeak = ST depression induced by exercise relative to rest
41. slope: the slope of the peak exercise ST segment

    -- Value 1: upsloping
    -- Value 2: flat
    -- Value 3: downsloping

42. rldv5: height at rest
43. rldv5e: height at peak exercise
44. ca: number of major vessels (0-3) colored by flourosopy
45. restckm: irrelevant
46. exerckm: irrelevant
47. restef: rest raidonuclid (sp?) ejection fraction
48. restwm: rest wall (sp?) motion abnormality

    0 = none
    1 = mild or moderate
    2 = moderate or severe
    3 = akinesis or dyskmem (sp?)

49. exeref: exercise radinalid (sp?) ejection fraction
50. exerwm: exercise wall (sp?) motion
51. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
52. thalsev: not used
53. thalpul: not used
54. earlobe: not used
55. cmo: month of cardiac cath (sp?) (perhaps "call")
56. cday: day of cardiac cath (sp?)
57. cyr: year of cardiac cath (sp?)
58. num: diagnosis of heart disease (angiographic disease status)

    -- Value 0: < 50% diameter narrowing
    -- Value 1: > 50% diameter narrowing
    (in any major vessel: attributes 59 through 68 are vessels)

59. lmt
60. ladprox
61. laddist
62. diag
63. cxmain
64. ramus
65. om1
66. om2
67. rcaprox
68. rcadist
69. lvx1: not used
70. lvx2: not used
71. lvx3: not used
72. lvx4: not used
73. lvf: not used
74. cathef: not used
75. junk: not used
76. name: last name of patient (I replaced this with the dummy string "name") .
The best 14 attributes are chosen for the Prediction.

Source:
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.
Further EDA and pre-processing needs to be carried out on the dataset to understand and evaluate it.

Machine Learning Methods

1. Logistic Regression (Scikit-learn)
2. Naive Bayes (Scikit-learn)
3. Support Vector Machine (Linear) (Scikit-learn)
4. K-Nearest Neighbours (Scikit-learn)
5. Decision Tree (Scikit-learn)
6. Random Forest (Scikit-learn)
7. XGBoost (Scikit-learn)
8. Artificial Neural Network with 1 Hidden layer (Keras)

The end result would be to compare the accuracy of these models to see how correctly they have classified the diagnosis of heart disease.

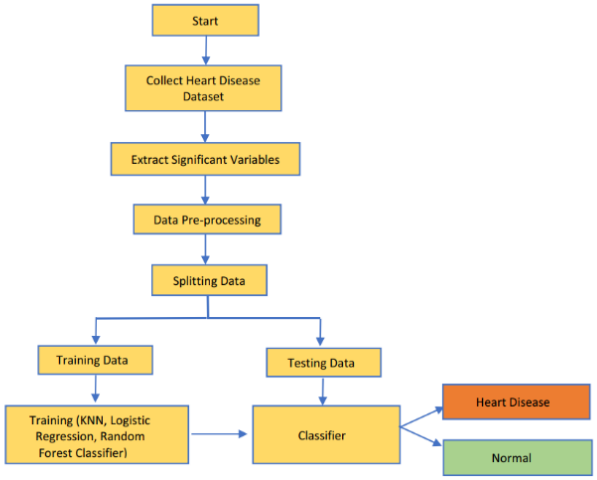
Assessment

Accuracy of the algorithms depends on four values namely true positive(TP), false positive(FP), true negative(TN) and false negative(FN).
Accuracy= (FN+TP) / (TP+FP+TN+FN)
The numerical value of TP, FP, TN, FN defines as:
TP= Number of persons with heart diseases
TN= Number of persons with heart diseases and no heart diseases
FP= Number of persons with no heart diseases
FN= Number of persons with no heart diseases and with heart diseases

Presentation and Visualization

A graphical representation to compare the relationship between age ranges and cardiovascular disease.
Further Catplot can be used to visualize categorical values like chest pain type CP, sex, fasting blood sugar FBS and target .
Whereas distplot can be used to visualize distributive variables like age, chol – cholestol level, trestbps – resting blood sugar, thalach – maximm heart rate etc, oldpeak.

Design of the proposed Model



Roles

Exploratory Data Analysis EDA and pre-processing will be done by the team together
Individual roles : The machine learning algorithms will be shared equally by the team members
Samiksha Rajbhushan Ullal : Logistic Regression, Support Vector Machine, Decision Tree, XGBoost
Karan R : Naïve Bayes, K- Nearest Neighbour, Random Forest, ANN with 1 Hidden layer
Comparison of accuracy metrics would be done together

Schedule

Date	Task to be completed
20/12/21	LA proposal
25/12/21	EDA and data preprocessing
05/01/22	Predicting using various machine learning models
15/01/22	Presentation
17/01/22	Report submission

Bibliography

A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

Sharma, Himanshu, and M. A. Rizvi. "Prediction of heart disease using machine learning algorithms: A survey." *International Journal on Recent and Innovation Trends in Computing and Communication* 5.8 (2017): 99-104.

A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, “Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools” International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.

Amandeep Kaur and Jyoti Arora,“Heart Diseases Prediction using Data Mining Techniques: A survey” International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.

Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Diseases Prediction”, 4th International Conference on Computing Communication And Automation(ICCCA), 2018.

M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” ProcediaTechnol., vol. 10, pp. 85–94, 2013.

Hazra, A., Mandal, S., Gupta, A. and Mukherjee, “ A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review” Advances in Computational Sciences and Technology , 2017.