Pune Institute of Computer Technology
Dhankawadi, Pune

A SEMINAR REPORT
ON

Machine Learning Algorithms used in Healthcare

SUBMITTED BY

Name
Roll No. 31217
Class TE2

Under the guidance of
Prof. A.A.Chandorkar

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2019-20

Pune Institute of Computer Technology
Dhankawadi, Pune-43

CERTIFICATE

This is to certify that the Seminar report entitled

“Machine Learning Algorithms used in Healthcare”

Submitted by
Sanket Rajesh Gattani Roll No 31217

has satisfactorily completed a seminar report under the guidance of Prof. A.A.Chandorkar towards the partial fulfillment of third year Computer Engineering Semester II, Academic Year 2019-20 of Savitribai Phule Pune University.

Prof. A.A.Chandorkar
Internal Guide

Prof. M.S.Takalikar
Head
Department of Computer Engineering

Place:
Date:

ACKNOWLEDGEMENT

I sincerely thank our Seminar Coordinator Prof. B.D.Zope and Head of Department Prof. M.S.Takalikar for their support.

I also sincerely convey my gratitude to my guide Prof. A.A.Chandorkar, Department of Computer Engineering for her constant support, providing all the help, motivation and encouragement from beginning till end to make this seminar a grand success.

Contents

List of Tables

List of Figures

Abstract

This seminar explains the use of Different Machine Learning Algorithms used in Healthcare. Machine Learning is highly used for health monitoring to reduce the mortality rate and enhance the life expectancy. Organs such as kidneys, liver, pancreas, eyes are highly affected in the run of life. Cancers like breast cancer have shown an increase in the count since last decade. This leads to invent new techniques in the field of medical sciences which can give accurate and timely predictions to reduce the mortality rate. Different Machine Learning models used by researchers will be explained and compared on the basis of quality and speed of algorithms.

Keywords

Disease Prediction, Machine Learning Algorithms, Classification Algorithms - Logistic Regression, KNN, SVM, Naive Bayes, Decision Tree, Random Forest, Precision and Accuracy.

1 INTRODUCTION

Machine learning in medicine has recently made headlines. Google has developed machine learning algorithms to help identify cancerous tumors. Stanford is using deep learning to identify skin cancer. A recent JAMA article reported the results of a deep machine-learning algorithm that was able to diagnose diabetic retinopathy in retinal images. It's clear that machine learning puts another arrow in the quiver of clinical decision making.

Machine learning Algorithms can provide immediate benefit to disciplines with processes that are reproducible or standardized. Consider the case of Heart Disease Prediction, it is dependent on features like gender, age, cigarettes per day, prevalent stroke, systolic blood pressure, glucose level. Many machine learning models like Logistic Regression, K-Nearest Neighbour, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest can be used for prediction. Efficiency, reliability, and accuracy will be the factors for the decision of the best fit algorithm.

Comparative study of different machine learning algorithms in prediction of various diseases and best working situations is taken into consideration during implementation of the sample model of "Diagnosis of Heart Disease". The outcome of this study will be categorization of machine learning algorithms as per situations taking their efficiencies and accuracy into consideration.

2 MOTIVATION

With growing populations and increased life expectancy, health systems are quickly becoming overburdened, under-resourced and not equipped for the challenges they face. Scientists have been working on ML models that predict disease susceptibility or aid in early diagnosis of diseases and illnesses. Many devices are using artificial intelligence algorithms for the precise detection of complex medical conditions in the field. It connects to existing medical sensors and can be used by non-doctor users to identify issues early, avoiding complications and hospitalizations. A new deep learning-based prediction model that can forecast the development of breast cancer up to five years in advance. Their model was trained on mammograms and patient follow-up data to identify patterns that would not be obvious to or even observable by human clinicians. The results have so far shown to be far more precise, especially at predictive, pre-diagnosis discovery.

Health Catalyst believes machine learning is the life-saving technology that will transform healthcare. This technology challenges the traditional, reactive approach to healthcare. In fact, it's the exact opposite: predictive, proactive, and preventative—life-saving qualities that make it a critically essential capability in every health system.

With the rise of machine learning and AI, There are various capacities where AI is emerging as a game-changer for the healthcare industry. Few examples in use today: Radiography, Patient risk identification, Drug discovery etc. Automation will help doctors a lot in timely predictions and cure.e

3 A SURVEY ON PAPERS

3.1 An Empirical Study of Machine Learning Algorithms for Cancer Identification

This paper talks about machine learning researchers that have recently proposed computational methods to improve the performance, and explains the need to apply these recent machine learning algorithms for clinical data generated via various technologies. In this paper, they used DeepBoost, which is a new ensemble learning algorithm; xgboost, which is a scalable end-to-end tree boosting system; a variant of Adaboost; and support vector machines (SVM). These machine learning algorithms are applied for cancer identification.

Outcome: Experimental results on real data pertaining to thyroid cancer, colon cancer, and liver cancer show that SVM outperforms the previously mentioned algorithms.

3.2 Prognosis of Diseases Using Machine Learning Algorithms

This paper explains how different machine learning algorithms are applied in disease prediction and its accuracy.

Outcome: Best fit algorithms for various diseases with accurate and precise predictions.

4 PROBLEM DEFINITION

Prediction of heart disease based on following features:

Gender

Age

Current Smoker

Cigarettes Per Day

BPMeds: 0 = Not on Blood Pressure medications; 1 = Is on Blood Pressure medications

PrevalentStroke

PrevalentHyp

Diabetes 0 = No; 1 = Yes

TotChol mg/dL

SysBPmmHg

DiaBPmmHg

BMIBody Mass Index calculated as: Weight (kg) / Height(meter-squared)

heartRateBeats/Min (Ventricular)

Glucose mg/dL

Prediction to be made:

Heart disease (0 = No; 1 = Yes)

5 DIFFERENT MACHINE LEARNING ALGORITHM

5.1 Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It works on sigmoid function It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc

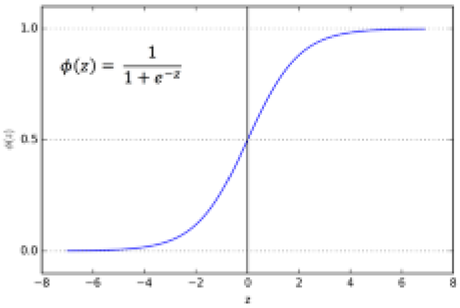


Figure 1: Logistic Regression Model

5.2 K-Nearest Neighbours

The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.Can be used for Regression as well as for Classification but mostly it is used for the Classification problems.K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data therefore it is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

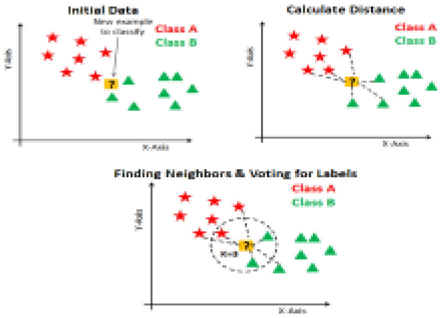


Figure 2: K-nearest Neighbour Model

5.3 Support Vector Machine

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.It is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces.

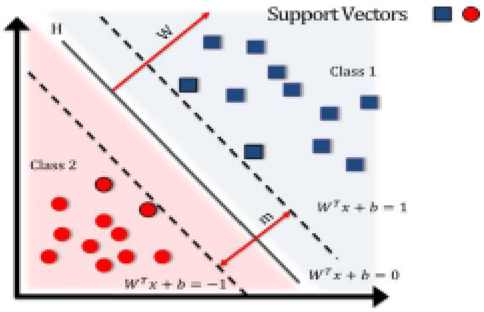


Figure 3: Support vector Machine Model

5.4 Naive Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:P(A—B) = P(B—A) P(A)/P(B)

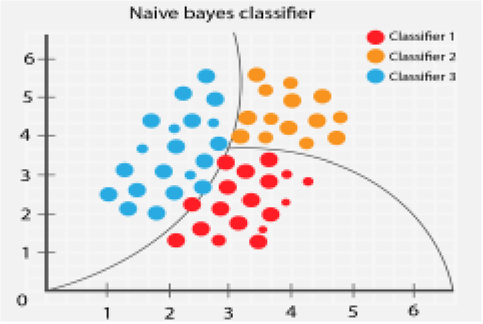


Figure 4: Naive Bayes Model

5.5 Decision Tree

Decision tree algorithms fall under the category of supervised learning. They can be used to solve both regression and classification problems. It uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can

represent any boolean function on discrete attributes using the decision tree.In Decision Tree the major challenge is to identify the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:
Information Gain: When we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information gain is a measure of this change in entropy.
Gini Index: Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.It means an attribute with lower Gini index should be preferred.Sklearn supports “Gini” criteria for Gini Index and by default, it takes “gini” value.

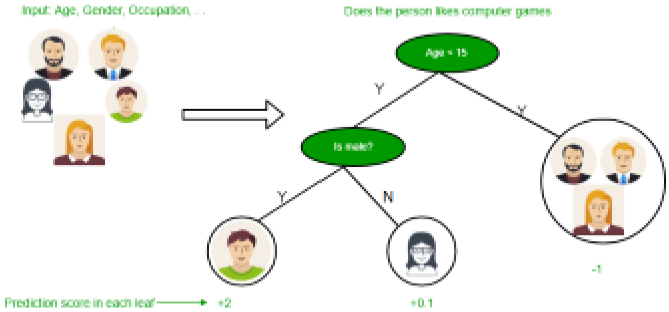

Input: Age, Gender, Occupation, . . .
Does the person likes computer games
Age < 15
Y
Y
Is male?
Y
N
-1
Prediction score in each leaf
+2
+0.1

Figure 5: Decision Tree Model

5.6 Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

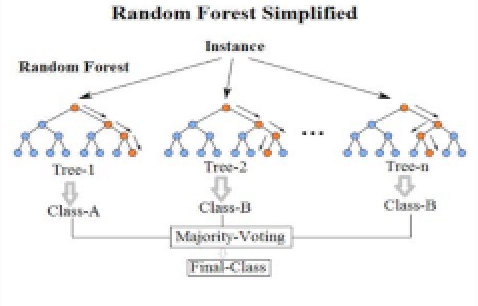


Figure 6: Random Forest Model

6 DISCUSSION ON IMPLEMENTATION AND RESULTS

6.1 Steps involved during implementation of machine learning Model

6.1.1 Data Collection

The quantity quality of your data dictate how accurate our model is. Dataset Used: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset

6.1.2 Data Preprocessing

This includes -
Wrangle data and prepare it for training
Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions,feature scaling etc.)
Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
Split into training and evaluation sets
After this step the dataset is scaled and clean.

6.1.3 Choose a model

Different algorithms are for different tasks; choose the right one

6.1.4 Train the Model

The goal of training is to answer a question or make a prediction correctly as often as possible
Linear regression example: algorithm would need to learn values for m (or W) and b (x is input, y is output)
Each iteration of process is a training step

6.1.5 Evaluate the Model

Uses some metric or combination of metrics to ”measure” objective performance of model
Test the model against previously unseen data
This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)
Good train/eval split?, depending on domain, data availability, dataset particulars, etc.

6.1.6 Parameter Tuning

This step refers to hyperparameter tuning, which is an ”artform” as opposed to a science
Tune model parameters for improved performance.
Simple model hyperparameters may include: number of training steps, learning rate, initialization values and distribution, etc.

6.1.7 Make Predictions

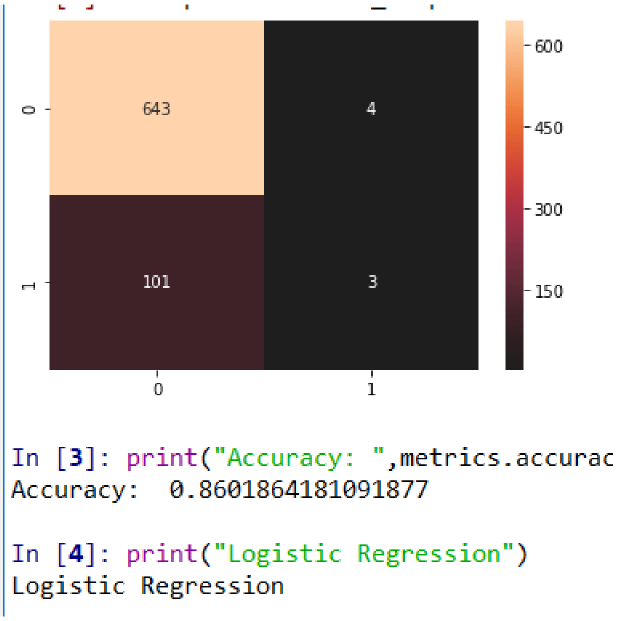
Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world

6.2 Implementation Results

Table 1: Machine Learning Algorithms and their accuracy in Heart Disease Prediction

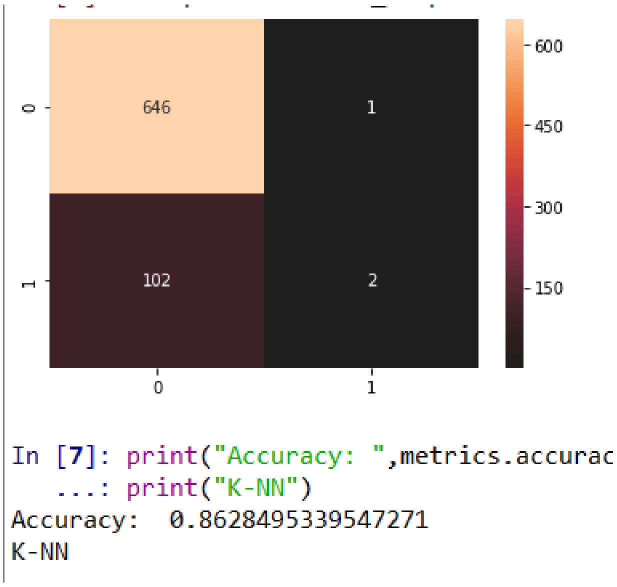
S.No	Algorithms	Accuracy
1	Logistic regression	0.86018
2	K-nearest Neighbour	0.86284
3	Support vector Machine	0.86018
4	Naive Bayes	0.85086
5	Decision Tree	0.76697
6	Random Forest	0.84420

6.3 HeatMaps For Results



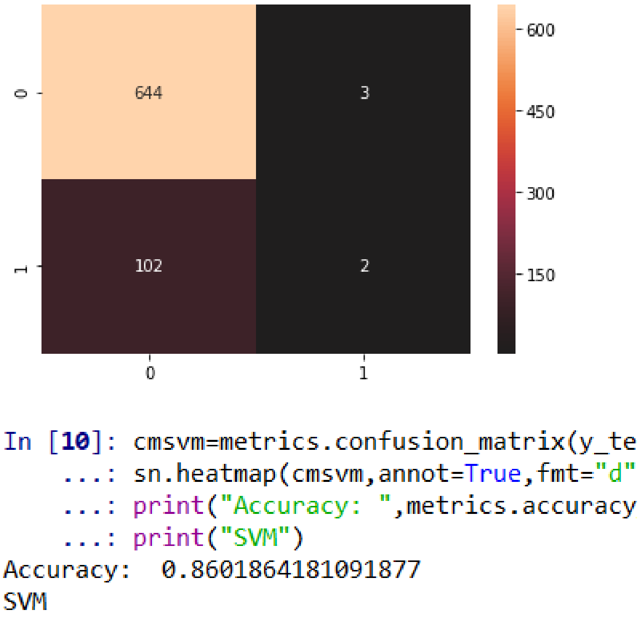
```
In [3]: print("Accuracy: ",metrics.accurac
Accuracy:  0.8601864181091877

In [4]: print("Logistic Regression")
Logistic Regression
```

Figure 7: Logistic Regression Model



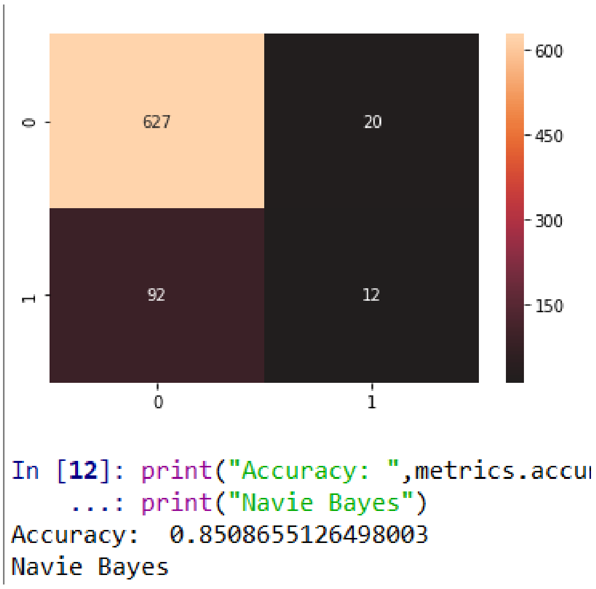
```
In [7]: print("Accuracy: ",metrics.accurac
   ...: print("K-NN")
Accuracy:  0.8628495339547271
K-NN
```

Figure 8: K-nearest Neighbour Model



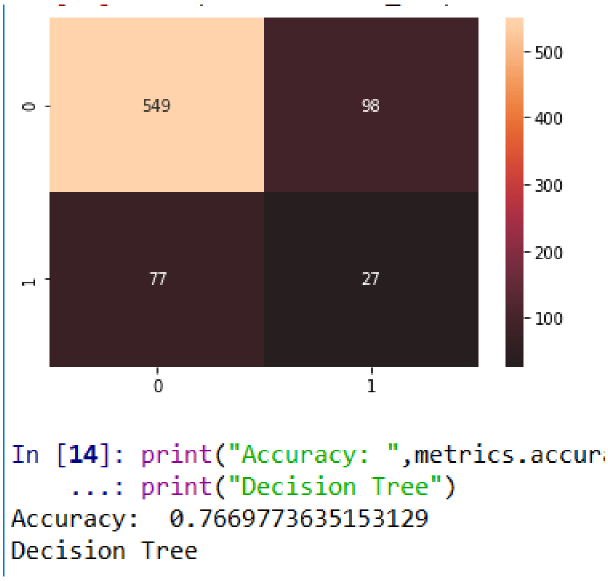
```
In [10]: cmsvm=metrics.confusion_matrix(y_te
    ...: sn.heatmap(cmsvm,annot=True,fmt="d"
    ...: print("Accuracy: ",metrics.accuracy
    ...: print("SVM")
Accuracy:  0.8601864181091877
SVM
```

Figure 9: Support vector Machine Model



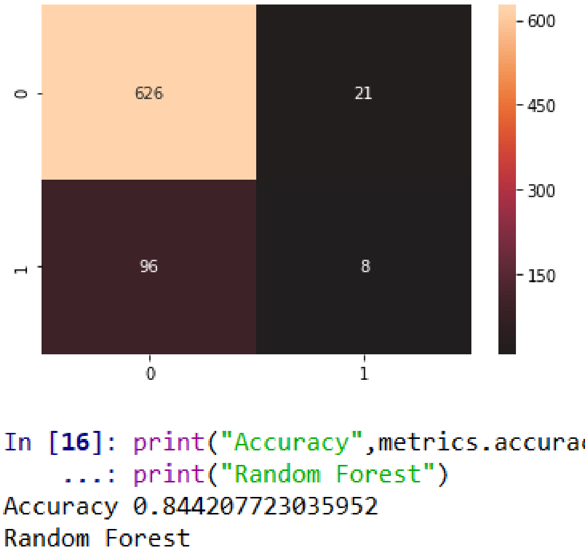
```
In [12]: print("Accuracy: ",metrics.accu
    ...: print("Navie Bayes")
Accuracy:  0.8508655126498003
Navie Bayes
```

Figure 10: Naive Bayes Model



```
In [14]: print("Accuracy: ",metrics.accur
    ...: print("Decision Tree")
Accuracy:  0.7669773635153129
Decision Tree
```

Figure 11: Decision Tree Model



```
In [16]: print("Accuracy",metrics.accura
    ...: print("Random Forest")
Accuracy 0.844207723035952
Random Forest
```

Figure 12: Random Forest Model

7 CONCLUSION

This Report summarizes the application of different machine learning algorithms for prognosis of the disease. In a study of different machine learning algorithms for heart disease prediction such as Logistic Regression, KNN, SVM, Naive Bayes, Decision Tree, Random Forest we come to know K-NN and SVM is working best with an accuracy of 0.86284 and 0.86018 respectively. SVM works best in case of higher dimensional visualizations and best fit extreme conditions whereas KNN works best if data is good enough to group items on basis of already plotted neighbours. Naive Bayes can be considered but Decision Tree and Random Forest do not work as good for predictions in healthcare.

References

[1] K. Shailaja, B. Seetharamulu, M. A. Jabbar, “Machine Learning in Healthcare: A Review”, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE,DOI: 10.1109/ICECA.2018.8474918

[2] Hanyue Dou, “Application of Machine Learning in The Field of Medical Science”, 2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE, DOI: 10.1109/YAC.2019.8787685

[3] Turki Turki, “An empirical study of machine learning algorithms for cancer identification”, 2018 IEEE International Conference on networking sensing and control (ICNSC), DOI: 10.1109/ICNSC.2018.8361312

[4] N. Marline Joys Kumari, Kishore k. v. krishna, “Prognosis of Diseases Using Machine Learning Algorithms”, 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), DOI: 10.1109/ICCTCT.2018.8550902

attach your review and visit log here......

attach plagiarism report here.....