Volume 5, Issue 11, November 2020 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Improving Accuracy of Classifier by Using Bagging Method of Random Forest

Rishika Patidar
P.G. Research Scholar
C.S.E. Department JIT Borawan
Khargone India

Mr. Abhay Mundra
Asst. Professor C.S.E. Department
JIT Borawan
Khargone India

Sachin Mahajan
Asst. Professor and H.O.D
C.S.E. Department
JIT Borawan Khargone India

Abstract: ***Random forest is a supervised learning algorithm which is used for both classification as well as regression. A forest is made up of trees and more trees mean more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests are a wonderful tool for making predictions considering they do not overfit because of the law of large numbers. Introducing the right kind of randomness makes them accurate classifiers and regressor . Random forest algorithm is considered as one of the most powerful algorithms because of its capability to find the relative importance of each feature/variable in the dataset. In this paper we create a model with the most important feature based on random forest with bagging method. The model is simpler and easy to interpret. Proposed reduces the variance and hence over-fitting and reduces the computational cost and time of training***

Keywords: Random forest ; Classifier; Learner; Tree; Over-fitting

I. INTRODUCTION

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. A forest is made up of trees and more trees mean more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Random Forests are an ensemble learning method (also thought of as a form of nearest neighbor predictor) for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.
Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests are a wonderful tool for making predictions considering they do not overfit because of the law of large numbers. Introducing the right kind of randomness makes them accurate classifiers and regressors.

Single decision trees often have high variance or high bias. Random Forests attempts to mitigate the problems of high variance and high bias by averaging to find a natural balance between the two extremes. Considering that Random Forests have few parameters to tune and can be used simply with default parameter settings, they are a simple tool to use without having a model or to produce a reasonable model fast and efficiently.

Random Forests are easy to learn and use for both professionals and lay people - with little research and programming required and may be used by folks without a strong statistical background. Simply put, you can safely make more accurate predictions without most basic mistakes common to other methods[10].

DIFFERENCE BETWEEN RANDOM FORESTS AND DECISION TREES

Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees. The following figure shows the decision boundary becomes more accurate and stable as more trees are added.

1. **Trees are unpruned**. While a single decision tree like CART is often pruned, a random forest tree is fully grown and unpruned, and so, naturally, the feature space is split into more and smaller regions[9,11,12].
2. **Trees are diverse**. Each random forest tree is learned on a random sample, and at each node, a random set of features are considered for splitting. Both mechanisms create diversity among the trees.
3. **Handling Over-fitting** A single decision tree needs pruning to avoid over-fitting. The following shows the decision boundary from an unpruned tree. The boundary is smoother but makes obvious mistakes (over-fitting).

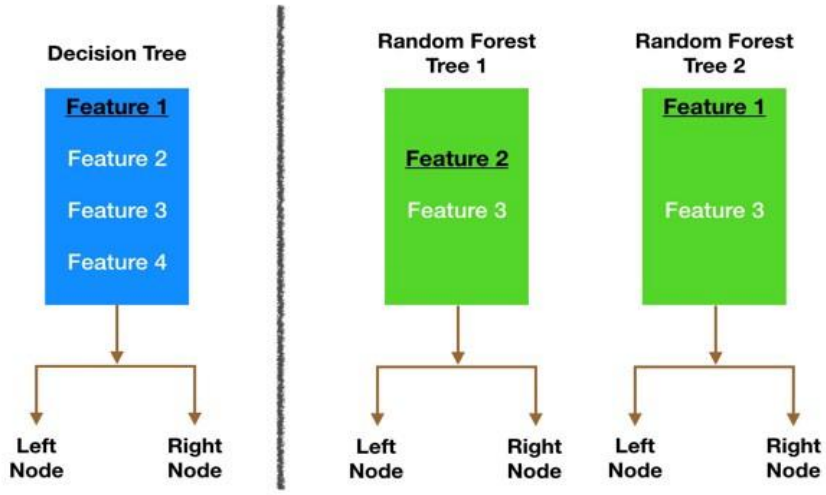

Figure 1 Difference between decision tree and random forest

PROPERTIES OF RANDOM FOREST

- Random forest is a predictive modeling algorithm (not any descriptive modeling algorithm).
- The random forest can be used for both classification and regression tasks.
- It works well with default hyper-parameters.
- It can be used to rank the importance of variables in a regression or classification problem.
- The correlation between any two trees in the forest. Increasing the correlation increases the forest error rate.
- A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate.
- It runs efficiently on large datasets.

II. LITERATURE SURVEY

In 2012 Gerard Biau "Analysis of a Random Forests Model". They offer an in-depth analysis of a random forests model suggested by Breiman , which is very close to the original algorithm. They showed in particular that the procedure is consistent and adapts to sparsity, in the sense that its rate of convergence depends only on the number of strong features and not on how many noise variables are present. The evolution of the MSE value with respect to *n* and *d*, for each model and the two tested procedures. They showed that the overall performance of the alternative method is very similar to the one of the original algorithm. They proposed a good approximation of the authentic Breiman's forests. They also showed that for a sufficiently large *n*, the capabilities of the forests are nearly independent of *d*, in accordance with the idea that the (asymptotic) rate of convergence of the method should only depend on the "true" dimensionality *S* . Finally, they showed that both algorithms perform well on the third model, which has been precisely designed for a tree-structured predictor[1].

In 2013 Dengju Yao Jing Yang proposed "A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines". They surveyed some kind of popular data mining techniques for disease prediction and diagnosis, such as decision tree, associated rule analysis and clustering analysis. Then they proposed a novel hybrid method of random forest and multivariate adaptive regression splines for building disease prediction model. Firstly, random forest algorithm is used to perform a preliminary screening of variables and to gain an importance ranks. They analyzed the characteristic of medical data and proposed a novel method of hybrid of RF and MARS for disease diagnosis and prediction. The proposed method is implemented on R software and is tested on the WDBC dataset. At the end, the performance of the hybrid algorithm of RF and MARS is compared with C4.5 algorithm and SVMMaj algorithm. The result experiment shows that the combination method of RF and MARS is suitable for disease prediction, which has not only good classification accuracy but will result in relatively simple and interpretable model [2].

In 2014 A.V. Lebedev,, E. Westman, G.J.P. Van Westenc "Random Forest ensembles for detection and prediction of Alzheimer3 disease with a good between-cohort robustness" They investigated the efficacy of Random Forest classifiers trained using different structural MRI measures, with and without neuron anatomical constraints in the detection and prediction of AD in terms of accuracy and between-

cohort robustness. From The ADNI database, 185 AD, and 225 healthy controls (HC) were randomly split into training and testing datasets. 165 subjects with mild cognitive impairment (MCI) were distributed according to the month of conversion to dementia (4-year follow-up). 1.5-TMRI-scans were processed using Free surfer segmentation and cortical reconstruction. Models3 between-cohort robustness was additionally assessed using the Add Neuro Med dataset acquired with harmonized clinical and imaging protocols. The Random Forest model resulted in significantly higher accuracy compared to the reference classifier (linear Support Vector Machine). The models trained using parcelled and high dimensional (HD) input demonstrated equivalent performance, but the former was more effective in terms of computation/memory and time costs. They applied to the independent AddNeuroMed cohort, the best ADNI models produced equivalent performance without substantial accuracy drop, suggesting good robustness sufficient for future clinical implementation [3].

In 2015 Prajwala T R "A Comparative Study on Decision Tree and Random Forest Using R Tool".Data mining is a process of extracting valuable information from large set databases. Classification a supervised technique is assigning data samples to target classes. They discussesed two classification algorithms namely decision trees and Random forest.. Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. Random forest includes construction of decision trees of the given training data and matching the test data with these. Rattle an open source R-GUI is used for analysis of weather data for prediction of rainfall using 256 data samples. Based on results obtained a comparative analysis is done. Classification in data mining assigns data samples to target classes. A random forest model is typically made up of tens or hundreds of decision trees. The proposed algorithm were compared and analysis was performed using the tool Rattle-R GUI , by considering 256 data samples of weather data set[4].

In 2016 Lakshmi Devasena C proposed " Proficiency Comparison of Random Forest and J48 Classifiers for Heart Disease Prediction". The term Heart disease covers the various diseases that affect the heart. The exposure of heart disease from various symptom or factors is an issue which is not free from false presumptions often accompanied by unpredictable effects. Identification of heart disease is a momentous and tedious task in medicine. It is requisite to find the best fit classification algorithm that has superior accuracy on classification in heart disease prediction. They compare the efficiency of Random forest and J48 classifiers for prediction of heart diseases using different measures. They investigated the efficiency of Random Forest and J48 Classifiers for heart disease prediction. Experimentation is accomplished using the open source machine learning tool. Effectiveness comparison of both the classifiers has been done using different scales of performance evaluation measures. Eventually, it is perceived that Random Forest Classifier performs better than J48 Classifier for heart disease prediction by taking measures including Classification accuracy, Errors and Time taken to build the model[5].

In 2017 Priya R. Patil S. A. Kinariwala propose " Automated Diagnosis of Heart Disease using Random Forest Algorithm". They a proposed decision support system made by three data mining techniques namely Classical Random Forest, Modified Random Forest and Weighted Random Forest. The classical random forests construct a collection of trees. In Modified Random Forest, the tree is constructed dynamically with online fitting procedure. A random forest is a substantial modification of bagging. Forest construction is based on three step process. Forest construction, The polynomial fitting procedure, The termination criterion , Weighted Random Forest, The Attribute Weighting Method is used for improving Accuracy of Modified Random Forest. There are Two Techniques are used in Attribute Weighting averaged One-Dependence Estimators (AODE) and decision Tree-based Attribute Weighted Averaged One-dependence Estimator (DTWAODE). They proposed an automated method for the determination of the number of base classifiers in the random forests classification algorithm using an online fitting procedure[6].

In 2018 H. Kaur, D. Gupta "Human Heart Disease Prediction System Using Random Forest Technique". Data mining is the analytical process to explore specific data from large volume of data. It is a process that finds previously unknown patterns and trends in databases. This information can be further used to build predictive models. Their main objective is to learn data mining techniques which can be used in the prediction of heart diseases using any data mining tool. Heart is the most vital part of the human body as human life depends upon efficient working of heart. A Heart disease is caused due to narrowing or blockage of coronary arteries. This is caused by the deposition of fat on the inner walls of the arteries and also due to build up cholesterol. Thus, a beneficial way to predict heart diseases in health care industry is an effective and efficient heart disease prediction system. This system will find human interpretable patterns and will determine trends in patient records to improve health care. They applied Random Forest technique to enhance the accuracy of the system. In the proposed work, heart disease prediction system was developed using classification algorithms through Matlab data mining tool to predict effective and accurate results regarding whether the patient is suffering from heart disease or not[7].

In 2019 B. Senthil Kumar, R. Gunavathi " AN Enhanced Model for Diabetes Prediction using Improved Firefly Feature Selection and Hybrid Random Forest Algorithm" Diabetes is a chronic disease that causes numerous amount of death each year. Untreated diabetes disturbs the proper functionality of other organs in human body. Hence early detection is a significant process to have a healthy life style. Usually the performance of the classification is affected due to the existence of high dimensionality in medical data.In this study a system model is proposed on Pima dataset to enhance the classification accuracy by eliminating the irrelevant features. Therefore it is important to choose a suitable feature selection approach that provides the better accuracy in disease prediction compared to prior study. Hence novel techniques Improved Firefly (IFF)and hybrid Random forest algorithm is proposed for feature selection and classification. The present study provides a better result with 96.3% accuracy. The efficiency of the present study is compared with the prior classification approaches. This study introduced a feature selection approach by applying the weight to the basic firefly algorithm[8].

In 2019 Senthilkumar Mohan, Chandrasegar Thirumalai, proposed "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making

decisions and predictions from the large quantity of data produced by the healthcare industry. They also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. They proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. They produced an enhanced performance level with an accuracy level of 88:7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM). The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease[9].

.

III. PROBLEM STATEMENT

The performance of classification techniques depends on the type of dataset that have taken for experiment. The main problem related to classification techniques are

1. Accuracy: - This includes accuracy of the classifier in term of predicting the class label, guessing value of predicted attributes.
2. Speed:-This include the required time to construct the model (training time) and time to use the model (classification/prediction time)
3. Robustness:-This is the ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.
4. Scalability:-Efficiency in term of database size.
5. Interpretability:-Understanding and insight provided by the model. Interpretability is subjective and therefore more difficult to assess

PROPOSED APPROACH

1. Randomly select **"k"** features from total **"m"** features.
   Where **k <= m**
2. Among the **"k"** features, calculate the node **"d"** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until "l" number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

The beginning of random forest algorithm starts with randomly selecting **"k"** features out of total **"m"** features. In the image, you can observe that we are randomly taking features and observations. In the next stage, we are using the randomly selected "k" features to find the root node by using the best split approach. The next stage, we will be calculating the daughter nodes using the same best split approach. Will the first 3 stages until we form the tree with a root node and having the target as the leaf node. Finally, we repeat 1 to 4 stages to create "n" randomly created trees. This randomly created tree forms the random forest**.**

OUTLINE OF PROPOSED APPROACH

First, start with the selection of random samples from a given dataset. Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree. In this step, voting will be performed for every predicted result. At last, select the most voted prediction result as the final prediction result.

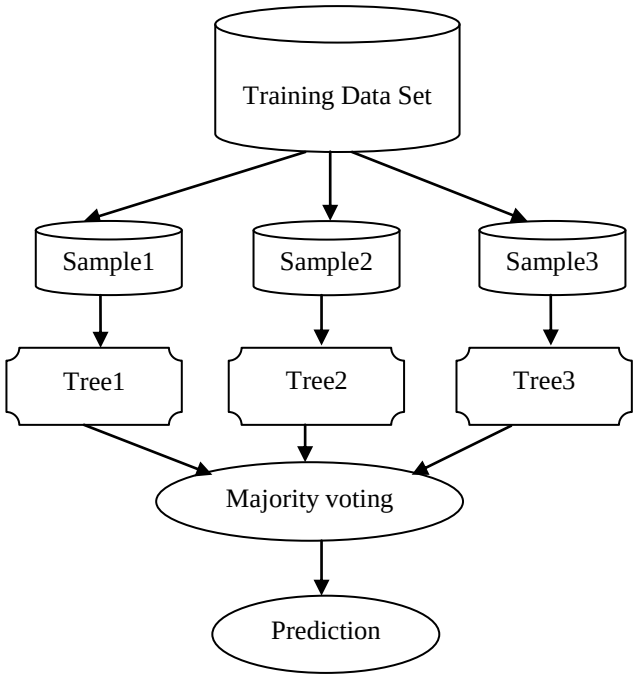

Figure 2 Outline and working of random forest

IV. IMPLEMENTATION ENVIRONMENT

We evaluate the performance of Random forest. We implement three tree (3 learners) form given data set. We divide the data set into three random parts based on the specified condition we create three tree. This form display data set used by tree1 and also implantation tree1. This tree started with blocked arties. The basic idea behind the tree1 is that is attribute blocked arties has the value no then we classified the tuple into no class. There is no need check others conditions .This tree started with chest pain. The basic idea behind the tree2 is that is attribute chest pain has the value no then we classified the tuple into no class. There is no need others conditions

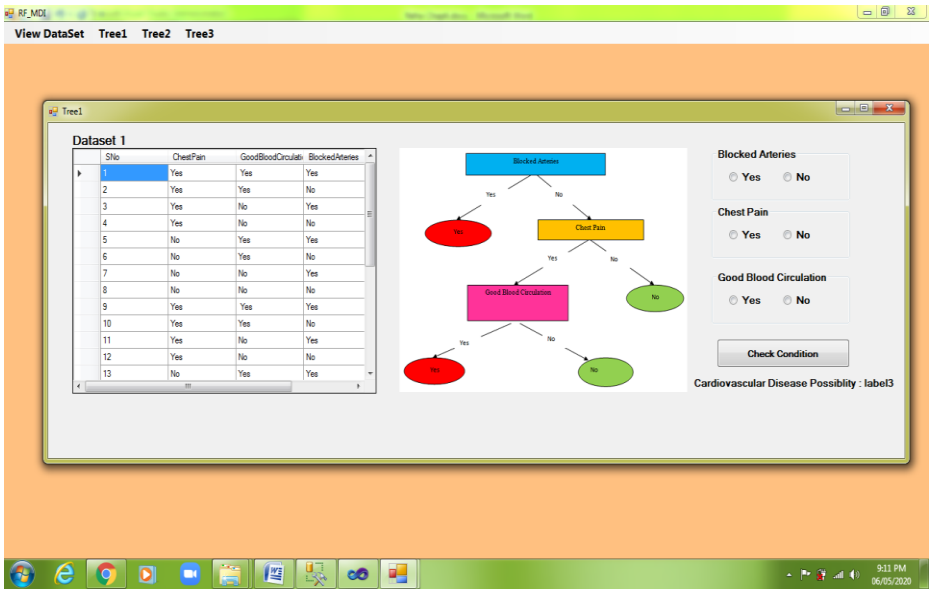

Figure 3 Implementation of proposed approach

COMPARATIVE ANALYSIS

Comparison based on no of records and classify by random forest into True Positive and Classify into True Negative. Out of 100 there are 80 records with yes class where tested records random forest classify 72 records into True Positive and 8 records into True Negative. Out of 200 there are 175 records with yes class where tested records random forest classify 162 records into True Positive and 13 records into True Negative. Out of 300 there are 265 records with yes class where tested records random forest classify 248 records into True Positive and 17 records into True Negative

Table 1 Number of records classify into True Positive and True Negative

Number of data records with Class Yes	Classify into True Positive	Classify into True Negative
80	72	8
175	162	13
265	248	17

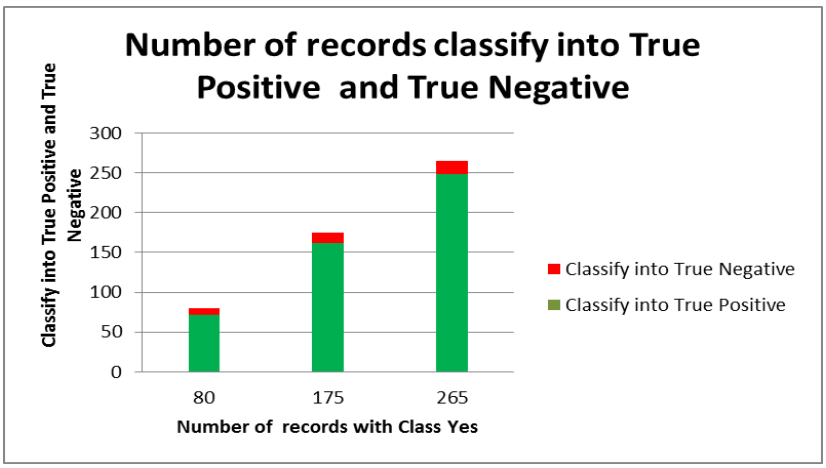

Figure 4 Number of records classify into True Positive and True Negative

CONCLUSION

Proposed model is simpler and easy to interpret. The proposed model reduces the variance and hence over-fitting. The Proposed model reduces the computational cost and time of training. In the proposed work we used loan data set we used bagging techniques of random forest to predict whether a person is classified into cardiovascular disease categories yes or not. In the proposed work we used heart patient data with some parameters Good Blood Circulation, Blocked Arteries etc . We create three trees (learner) for random forest. With the experimental analysis it is found that the proposed approach classifiers the data more accurately as compared to other classifiers.

REFERENCE

1. Gerard Biau "Analysis of a Random Forests Model" Journal of Machine Learning Research 13 (2012) 1063-1095 Submitted 10/10; Revised 10/11; Published 4/12
2. Dengju Yao and Jing Yang "A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines" JOURNAL OF COMPUTERS, VOL. 8, NO. 1, JANUARY 2013
3. A.V. Lebedeva,□, E.Westmanb "Random Forest ensembles for detection and prediction of Alzheimer3s disease with a good between-cohort robustness" NeuroImage: Clinical 6 (2014) 115–125 Stavanger University Hospital, PO Box 8100, 4068 Stavanger, Norway.
4. Prajwala T R "A Comparative Study on Decision Tree and Random Forest Using R Tool" International Journal of Advanced Research in Computer and Communication EngineeringVol. 4, Issue 1, January 2015
5. Lakshmi Devasena C" Proficiency Comparison of Random Forest and J48 Classifiers for Heart Disease Prediction" International Journal of Computing Academic Research (IJCAR) ISSN 2305-9184, Volume 5, Number 1 (February 2016), pp.46-55 © MEACSE Publications
6. Priya R Patil S. A. Kinariwala " Automated Diagnosis of Heart Disease using Random Forest AlgorithmInternational Journal of Advance Research, Ideas and Innovations in Technology ISSN: 2454-132X Impact factor: 4.295 (Volume3, Issue2)
7. H. Kaur1*, D. Gupta2 " Human Heart Disease Prediction System Using Random Forest Technique" International Journal of Computer Sciences and Engineering Open Access Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693
8. B. Senthil Kumar, R. Gunavathi" AN Enhanced Model for Diabetes Prediction using Improved Firefly Feature Selection and Hybrid Random Forest Algorithm" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019
9. Senthilkumar Mohan , Chandrasegar Thirumalai, Effective Heart Disease Prediction Using Hybrid Machine Learning TechniquesReceived May 13, 2019, accepted June 9, 2019, date of publication June 19, 2019, date of current version July 3, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2923707
10. Indu Yekkala Sunanda Dixit "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection" International Journal of Big Data and Analytics in Healthcare Volume 3 • Issue 1 • January-June 2018
11. Shanmuga Priya.Sbinaya.M Feature Selection using Random Forest Technique for the prediction of pest attack in cotton crops. International Journal of Pure and Applied Mathematics Volume 118 No. 18 2018, 2899-2903 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version) url: http://www.ijpam.eu
12. Youness Khourdifi1* Mohamed Bahaj1" Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization" International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019 DOI: 10.22266/ijies2019.0228.24