COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS WITH DIFFERENT CLASSIFIERS

[1]Anu C.S, [2]Preethi B, [3]Dr. Nirmala C.R
[1,2]Assistant Professor, CS&E Dept. B.I.E.T College, Davangere
[3]Head of Dept., CS&E Dept. B.I.E.T College, Davangere

ABSTRACT—
Healthcare practices include collecting all kinds of patient data which would help the doctor correctly diagnose the health condition of the patient. These data could be simple symptoms observed by the subject, initial diagnosis by a physician or a detailed test result from a laboratory. Thus, these data are only utilized for analysis by a doctor who then ascertains the disease using his/her personal medical expertise. Machine learning calculations can make sense of how to perform imperative errands by summing up from illustrations. This research aims at comparing different algorithms used in machine learning. Machine Learning can be both experience and explanation-based learning. In this study most popular algorithms were used like Decision Tree(DT), Random Forest (RF) and Gaussian Naïve Bayes(NB) and heart disease dataset have been used to check the efficiency of algorithms. Comparative analysis of the classifiers shows that Random Forest out performs the other methods with a high accuracy. The artificial intelligence has been used with Naive Bayes classification, Decision Tree and Random Forest classification algorithm to classify heart disease to check whether the patient is affected by disease or not. A performance analysis of the disease data among three algorithms is calculated and compared. The results of the simulations show the effectiveness of the classification techniques on a dataset, as well as the nature and complexity of the dataset used.

Keywords
Decision Tree, Random Forest, Gaussian Naïve Bayes, UCI dataset

1. INTRODUCTION

Due to modern lifestyle, diseases are increasing rapidly. Our lifestyle and food habit leads to create impact on our health causing heart diseases and other health issues. The amount of data in the medical industry is increasing day by day. It is a challenging task to handle a large amount of data and extracting productive information for effective decision making. For this reason, medical industry demands to apply a special technique. Machine learning (ML) is categorized under artificial intelligence of (AI) which facilitates the computer with efficiency to perform and learn even after not being particularly programmed. ML is a strategy for information examination that robotizes logical model building [1-3].ML only concentrates on developing computer programs flexible to change whenever expose to new data. Different ML algorithms involve huge potential to be successfully applied in different fields like medicals [1-5], corporates, education, robotics, games and much more [6]. In ML one of the important factors is to make machines able to learn efficiently and effectively [7]. There are an extensive variety of computations which help in making gadgets and strategy in ML. Sometimes these methods create confusion in their applicability in suitable methods and which algorithms gives more accuracy can't know. Researchers have used different algorithms in ML according to expertise, availability and the dataset [8]. Although ML is a discreetly young ground of research [9]. Selecting algorithm in ML for the given datasets (problem) can be tricky. Preprocessing strategies have been utilized to get highlights from bigger informational indexes to prepare the current classifier systems [10]. A comparative analysis is put together in investigating the improved accuracy of

classifiers [11]. The objective of these exercises isn't to include new usefulness, however to show the best strategy in examination with these DT,RF,NB strategies. Cross validation is used with 90% for Train and 10% for Test the dataset. Also find the accuracy, specificity and sensitivity for comparing ML algorithms.

2. LITERATURE SURVEY

Mohammed Abdul Khaleel has given paper in the Survey of Techniques for mining of data on Medical Data for Finding Frequent Diseases locally. This paper focus on dissect information mining procedures which are required for medicinal information mining particularly to find locally visit illnesses, for example, heart infirmities, lung malignancy, bosom disease et cetera. Information mining is the way toward extricating information for finding inactive examples which Vembandasamy et al. performed a work, to analyze and detect heart disease. In this the algorithm used was Naive Bayes algorithm. In Naïve Bayes algorithm they used Bayes theorem. Hence Naive Bayes has a very power to make assumption independently. The used data-set is obtained from a diabetic research institutes of Chennai, Tamilnadu which is leading institute. There are more than 500 patients in the dataset. The tool used is Weka and classification is executed by using 70% of Percentage Split. The accuracy offered by Naive Bayes is 86.419%. [12]

Costas Sideris, Nabil Alshurafa, Haik Kalantarian and Mohammad Pourhomayoun have given a paper named Remote Health Monitoring Outcome Success prediction using First Month and Baseline Intervention Data. RHS systems are effective in saving costs and reducing illness. In this paper, they portray an upgraded RHM framework, Wanda- CVD that is cell phone based and intended to give remote instructing and social help to members. CVD counteractive action measures are perceived as a basic focus by social insurance associations around the world.[13]

L.Sathish Kumar and A. Padmapriya has given a paper named Prediction for similarities of disease by using ID3 algorithm in television and mobile phone. This paper gives a programmed and concealed way to deal with recognize designs that are covered up of coronary illness. The given framework utilize information mining methods, for example, ID3 algorithm. This proposed method helps the people not only to know about the diseases but it can also help's to reduce the death rate and count of disease affected people.[14]

M.A.Nishara Banu and B.Gomathy has given a paper named Disease Predicting system using data mining techniques. In this paper they talk about MAFIA (Maximal Frequent Item set algorithm) and K-Means clustering. As classification is important for prediction of a disease. The classification based on MAFIA and K-Means results in accuracy.[15]

Wiharto and Hari Kusnanto have given a paper named Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm. In this paper they exhibit an expectation framework for heart infection utilizing Learning vector Quantization neural system calculation The neural system in this framework acknowledges 13 clinical includes as information and predicts that there is a nearness or nonattendance of coronary illness in the patient, alongside various execution measures. [16]

D.R. PatiI and Jayshril S. Sonawane have given a paper named Prediction of Heart Disease Using Learning Vector Quantization Algorithm. In this paper they exhibit an expectation framework for heart infection utilizing Learning vector Quantization neural system calculation The neural system in this framework acknowledges 13 clinical includes as information and pre- dicts that there is a nearness or nonattendance of coronary illness in the patient, alongside various execution measures.[17]

3. THE USED CLASSIFIERS

Decision Tree (DT)

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

$$Info(D) = -\sum_{i=1} p_i \, log_2(p_i)$$

D – Current state,

Pi – Probability of an event i of state D or Percentage of class I in a node of state D. There are two main types of Decision Trees:

1. Classification trees (Yes/No types): What we've seen above is an example of classification tree, where the outcome was a variable like 'fit' or 'unfit'. Here the decision variable is Categorical.

Classification : G = sum(pk * (1 — pk))

Pk - Proportion of same class inputs present in a particular group

2. Regression trees (Continuous data types): Here the decision or the outcome variable is Continuous, e.g. a number like 123. Working Now that we know what a Decision Tree is, we'll see how it works internally. There are many algorithms out there which construct Decision Trees, but one of the best is called as ID3 Algorithm. ID3 Stands for Iterative Dichotomiser 3. Before discussing the ID3 algorithm, we'll go through few definitions. Entropy, also called as Shannon Entropy is denoted by H(S) for a finite set S, is the measure of the amount of uncertainty or randomness in data.
Regression : sum(y — y')

Random Forest (RF)

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Applications of Random Forest

- Banking: Banking sector mostly uses this algorithm for the identification of loan risk.
- Medicine: With the help of this algorithm, disease trends and risks of the disease can be identified.
- Land Use: We can identify the areas of similar land use by this algorithm.
- Marketing: Marketing trends can be identified using this algorithm.

Advantages of Random Forest

- For applications in classification problems, Random Forest algorithm will avoid the overfitting problem.
- For both classification and regression task, the same random forest algorithm can be used .
- The Random Forest algorithm can be used for identifying the most important features from the training dataset, in other words, feature engineering.

Naïve Bayes (NB)

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

$$P(A \mid B) = P(B \mid A)\, P(A) / P(B)$$

Where,
P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

In the present study two different datasets are used. First data is used quality of red and white wine with 6500 rows and 13 columns [21]. The second one is biodegradable chemical with1055 rows and 13 columns [22]. All information

procured from UCI machine learning storehouse.

The methodology for the classification of these datasets is displayed in Figure 1. The analysis has been performed in Jupyter Notebook IDE running on (Intel i5 processor) with 4 GB RAM installed. The different data sets are taken as input for feature extractor and classification algorithm. The datasets are passed through a sequence of pre-processing blocks.

4. METHODOLOGY

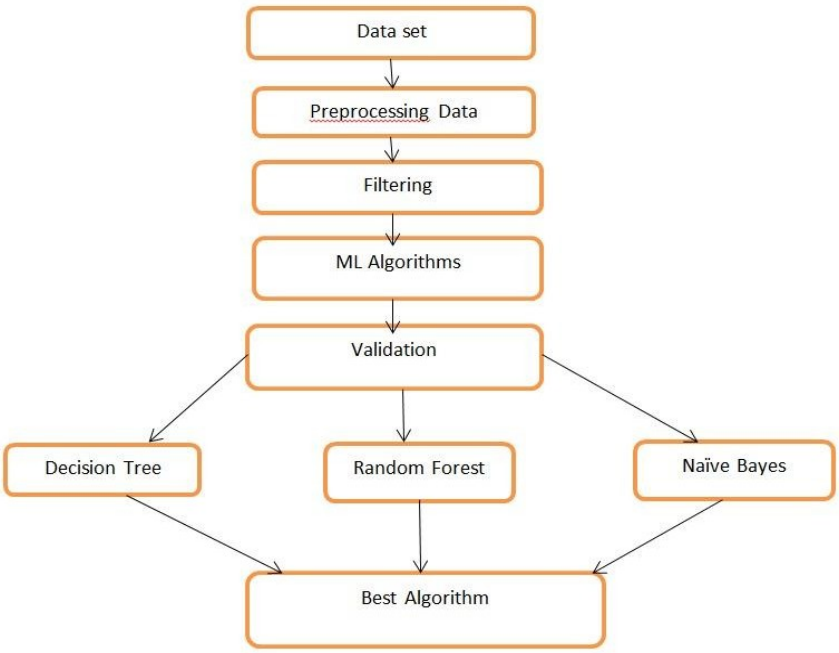

Fig : BLOCK Diagram of different classifiers

Dataset:

In the current study we have used dataset . This dataset were downloaded from ML repository UCI. Which is of heart disease dataset.

Pre-processing:

After the gathering of data next phase is to perform the preprocessing on the collected data. It is the technique that changes the raw data to the understandable format. If there are any existing null values, by this process they will be removed from the dataset.

Filtering:

After preprocessing of data next phase is to filter the data. After removing null values we will get the output of balanced dataset.

ML Algorithms:

We used different algorithms such as Decision Tree, Random Forest and Naive Bayes.

Validation:

In the validation 10-fold cross-validation method is used. To create the training and testing sets, all normalized features are randomized before they can be used to train the classifier networks; RF, DT, and NB. The separation is computed between test information and every case of preparing information. This separation controlled by various highlights utilized for the grouping. In this we have random forest classifier which given the higher accuracy among the three classifiers.

5. Results and Discussion

	Decision Tree	Random Forest	Gaussian Naïve Bayes	Best Score
Accuracy	0.534990	0.633770	0.597891	Random Forest
Precision	0.538603	0.638500	0.717736	Gaussian Naïve Bayes
Recall	0.488401	0.621091	0.323214	Random Forest
F1 Score	0.511910	0.629429	0.441818	Random Forest

Table1: Accuracy, Precision, Recall, F1 Score of Navie Bayes, Random Forest, Decision Tree

- **Accuracy Calculation**

The prediction accuracy is calculated using the formulae

Accuracy = (AP + AN) / (M + N)
Where, M = AP + AN and N = ALP + AN. Or AP + AN (TOTAL)

- **Precision (positive predictive value)**

Precision (PREC) is a classification technique which is used to find the items that are incorrectly labeled among the given class. The best precision result is 1.0, whereas the worst
one is 0.0.

PREC = TP / (TP + FP)

- **Recall**

Sensitivity (SN) is calculated using the number of correct positive prediction value divided by the total number of positive predictions. It is also called as recall (REC) or true positive rate (TPR). The best value is 1.0 and the worst value is 0.0.
SN = TP / (TP + FN) / (TP / P)

- **F1-score**

F1-score is a weighted average of recall and calculated precision value.
F1 = 2TP / (2TP + FP + FN)

A1 male

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
2	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
3	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
4	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
5	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
6	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
7	0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
8	0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
9	0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
10	1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0
11	1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0
12	0	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0
13	0	43	2	0	0	0	0	0	0	247	131	88	27.64	72	61	0
14	1	46	1	1	15	0	0	1	0	294	142	94	26.31	98	64	0
15	0	41	3	0	0	1	0	1	0	332	124	88	31.31	65	84	0
16	0	39	2	1	9	0	0	0	0	226	114	64	22.35	85	NA	0
17	0	38	2	1	20	0	0	1	0	221	140	90	21.35	95	70	1
18	1	48	3	1	10	0	0	1	0	232	138	90	22.37	64	72	0
19	0	46	2	1	20	0	0	0	0	291	112	78	23.38	80	89	1
20	0	38	2	1	5	0	0	0	0	195	122	84.5	23.24	75	78	0

Fig1: Dataset of heart disease

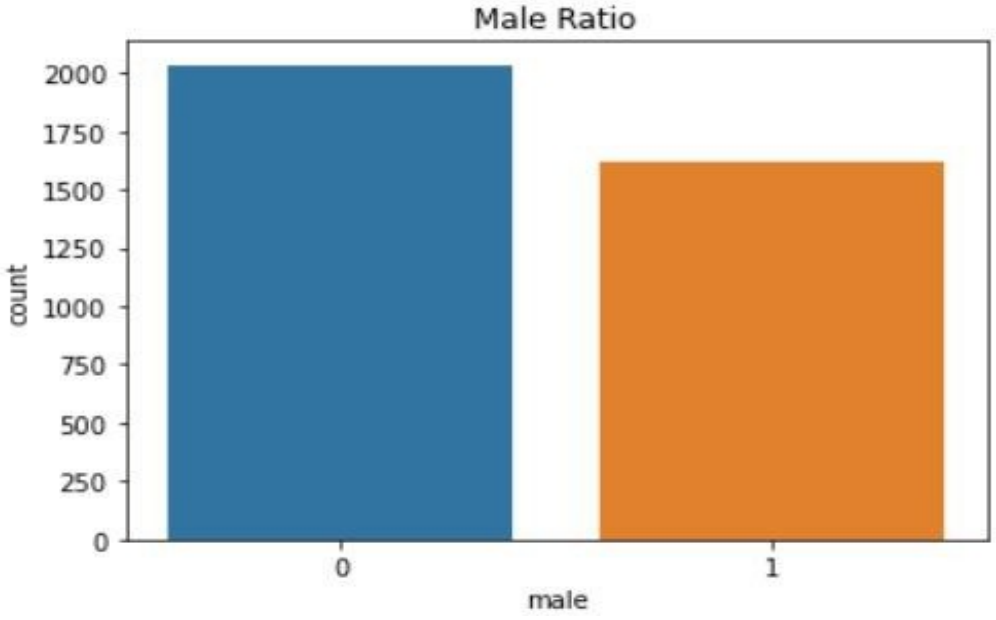

Fig2: Comparision of Male Count Ratio

According to the plot above, the Heart Study contains more data points corresponding to men.

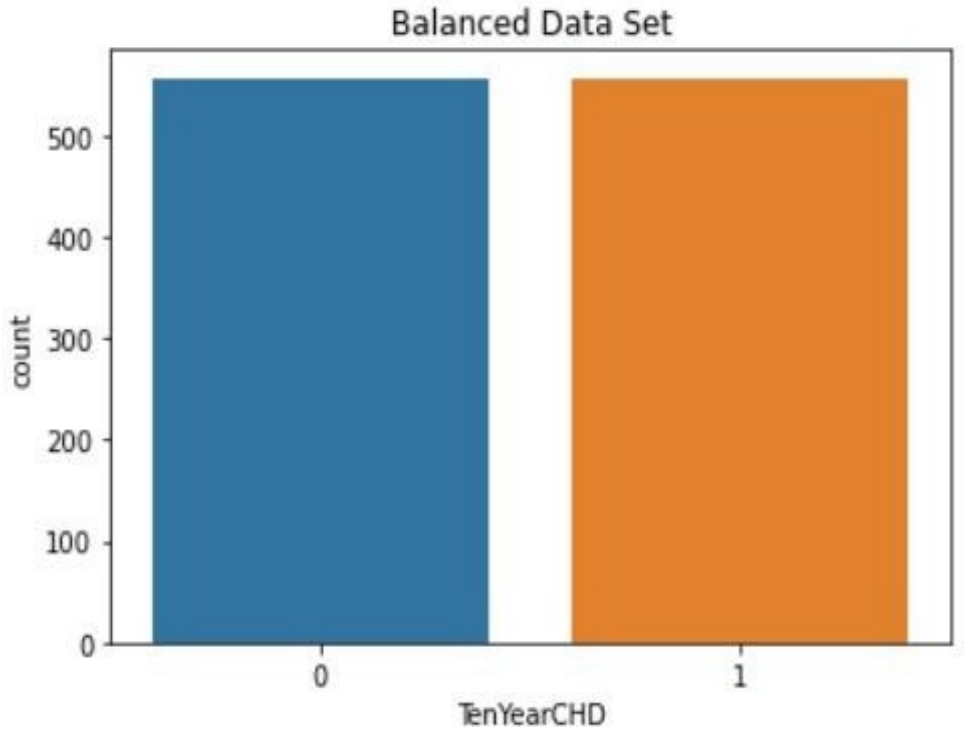

Fig3: Balanced Data Set of Ten Year Child

The plot above shows an equal number of classes is equal after having used the Random Under Sampling technique to balance the data set.

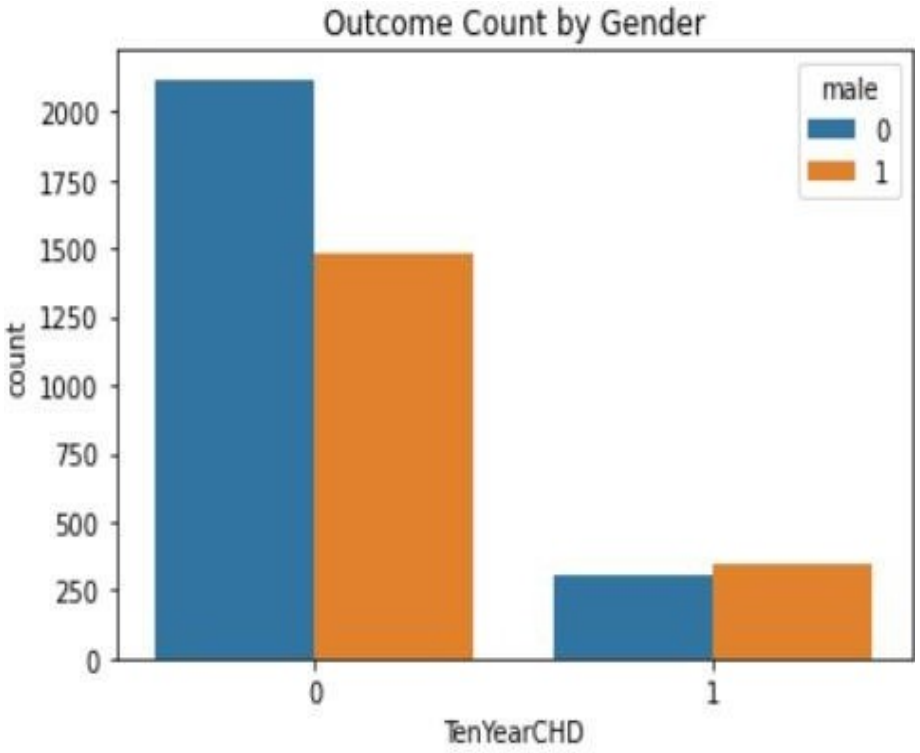

Fig4: Comparison of Outcome Count by Gender

Even though the total number of data points corresponding to men was lower, the plot above suggests that the risk of developing a heart disease on men is higher than on women.

6. CONCLUSION

As per current examination the relative investigation of DT, RF, and NB has been executed and the execution metric mirrors the execution of RF better when contrasted with the other broke down classifier. The dataset is chosen from online repositories. The techniques of pre- processing applied are filled in missing values and removing correlated columns. Next, the classifier is applied to the preprocessed dataset, and then Bayesian and random forest models are constructed. Finally, the accuracy of the models is calculated and analyses are based on the efficiency calculations. Bayesian Classification network shows the accuracy of 82.35 % for heart disease. Similarly, classification with Random forest model shows the accuracy ofwell compared with other two algorithms. When performing classification in the trained model by applying sample test data of disease, the random forest model gives accurate results. The proposed model works well against train data and test data further this model will provide the better results for real-time data.

7. REFERENCES

[1] Sharma, L., Gupta, G. and Jaiswal, V., 2016, December. Classification and development of tool for heart diseases (MRI images) using machine learning. In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on (pp. 219-224). IEEE.

[2] Chauhan, D. and Jaiswal, V., 2016, October. An efficient data mining classification approach for detecting lung cancer disease. In Communication and Electronics Systems (ICCES), International Conference on (pp. 1-8). IEEE.

[3] Negi, A. and Jaiswal, V., 2016, December. A first attempt to develop a diabetes prediction method based on different global datasets. In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on (pp. 237-241). IEEE.

[4] Pal, T., Jaiswal, V. and Chauhan, R.S., 2016. DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. Computers in biology and medicine, 78, pp.42-48.

[5] Jaiswal, V., et al., Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. BMC bioinformatics, 2013. 14(1): p. 211.

[6] Jaiswal, V., Chanumolu, S.K., Gupta, A., Chauhan, R.S. and Rout, C., 2013. Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. BMC bioinformatics, 14(1), p.211.

[7] Das, S., Dey, A., Pal, A. and Roy, N., 2015. Applications of Artificial Intelligence in Machine Learning: Review and Prospect. International Journal of Computer Applications, 115(9).

[8] Sebastiani, F., 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), pp.1-47.

Circulation in Computer Science International Conference on Innovations in Computing (ICIC 2017), pp:87-91 www.ccsarchive.org

[9] Cunningham, S.J., Littin, J. and Witten, I.H., 1997. Applications of machine learning in information retrieval.

[10] Mitchell, T.M., 2006. The discipline of machine learning (Vol. 3). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[11] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F. and Dennison, D., 2015. Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems (pp. 2503- 2511).

[12] Portugal, I., Alencar, P. and Cowan, D., 2015. The use of machine learning algorithms in recommender systems: a systematic review. arXiv preprint arXiv:1511.05263.

[13] Cost, S. and Salzberg, S., 1993. A weighted nearest neighbor algorithm for learning with symbolic features. Machine learning, 10(1), pp.57-78.

[14] Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M. and Scarpetta, S., 2003. Discrimination of earthquakes and underwater explosions using neural networks. Bulletin of the Seismological Society of America, 93(1), pp.215-223.

[15] Shalev-Shwartz, S. and Ben-David, S., 2014. Understanding machine learning: From theory to algorithms. Cambridge university press.

[16] Wagstaff, K., 2012. Machine learning that matters. arXiv preprint arXiv:1206.4656.

[17] Bennett, K.P. and Parrado-Hernández, E., 2006. The interplay of optimization and machine learning research. Journal of Machine Learning Research, 7(Jul), pp.1265- 1281.

[18] Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.

[19] Javidi, B., 2002. Image recognition and classification: algorithms, systems, and applications. CRC Press.

[20] Domingos, P., 2012. A few useful things to know about machine learning. Communications of the ACM, 55(10), pp.78-87.

[21] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), pp.547-553.

[22] Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R. and Consonni, V., 2013. Quantitative structure–activity relationship models for ready biodegradability of chemicals. Journal of chemical information and modeling, 53(4), pp.867-878.

[23] Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861- 874.