Proceedings of the 7th International Conference on Intelligent Computing and Control Systems (ICICCS-2023)
IEEE Xplore Part Number: CFP23K74-ART; ISBN: 979-8-3503-9725-3

2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) | 979-8-3503-9725-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICCS56967.2023.10142906

Comparative Analysis of Different Machine Learning Algorithms on Different Datasets

DR.G. S Bapi Raju
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India gsbapiraju@gmail.com

Chintala Manasa
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India chintalamanasa2001@gmail.com

Nandikatti Durga Bhavani
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India ndbhavani8@gmail

Jadi Amulya
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India annujadi@gmail.com

Dangatla Shirisha
Computer Science and Engineering Department
Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India dangatla.shirisha@gmail.com

Abstract -- Machine Learning is used to train models and machines without the help of any human interventions and guides. Here the models and machines are trained using algorithms. As it is difficult to train the algorithms one by one and get the analysis of the best algorithm. This research study intends to compare different algorithms used in different datasets to know the best algorithm according to different domains and the overall best algorithm. In this study, most popular supervised algorithms like SVM, Decision Tree, Random Forest, KNN, Logistic Regression, XGBoost, Adaboost, Voting Classifier (Decision Tree + Random Forest), Bagging Classifier (KNN + SVM) and Naïve Bayes. This research study has chosen three different domains for datasets. Different datasets were used to check the efficiency of algorithms. Comparative analysis of the classifiers shows that all algorithms outperform the existing methods with a high accuracy.

Keywords -- SVM, Decision Tree, Random Tree, KNN, Logistic Regression, XGBoost, Adaboost, Voting Classifier, Bagging Classifier, Naive Bayes, Confusion Matrix, Precision, Recall, Accuracy

1. INTRODUCTION

Machine Learning (ML) is the study of computer programs that use algorithms and statistical models to learn through inference and patterns without being explicitly programmed. ML algorithms find techniques, train models, and use the learned approach to determine the output automatically. A model is a machine learning system that has been trained to identify types of patterns using an algorithm.

This paper compares different machine learning algorithms using different Datasets.

ALGORITHMS: Support Vector Machine, K Nearest Neighbors (KNN) Classifier, Random Forest Classifier, Decision Tree Classifier, Logistic Regression, XGBoost, Voting Regression, Voting Classifier, AdaBoost Classifier, Bagging Classifier, Naive Bayes.

DATASETS: Health Care - Heart Disease: In this dataset, there are many attributes which cause heart disease to a person. So based on the data we can predict the chances of getting a heart disease for a person.

Agriculture - Crop Yield Production: In this crop yield dataset there are attributes related to region, area, season, crop type and amount of yield produced all over India. By making use of the above data we can predict how much yield can be produced under related circumstances.

Education - Student Placement Performance: Here, the dataset contains attributes related to board of education, marks obtained, whether placed or unplaced, if placed then the amount of salary earned. By utilizing the above information we can predict whether the student can get placed or not, if placed then we can say how much salary he/she can expect.

2. LITERATURE REVIEW

In this study of algorithms, Machine Learning has been used to compare its algorithms in many different ways. The works listed below are all relevant to this study.

The Authors of the paper [1] employed three models—neural networks, k-nearest neighbours, and support vector machines—on two datasets to evaluate the performance of the algorithms. The classifier analysis shows that SVM performs more accurately and efficiently than the other techniques.

The Authors of the paper [2] compared the best machine learning model using the performance parameters R squared and Mean Square Error. This research investigates which method outperforms the others when a comparison analysis is performed to determine the optimal model.

The Authors of the paper [3] have used 3 models like k-Nearest Neighbors, Support Vector Machine, Neural Network.This paper main aim is to find a best method to deal with classification problems of different datasets.

The Authors of the paper [4] have used 4 models like Support Vector Machine, Logistic Regression, Decision Tree, Random Forest. The main aim is to recognize Dementia among various patients. The system is simple and can easily help people by detecting Dementia among them.

3. ALGORITHMS

This section offers a concise explanation of all the supervised algorithms used in the proposed work.

3.1 Support Vector Machine

Both classification plus regression were executed utilizing a robust and versatile machine learning approach referred to as support vector machines discovering a hyper-plane in N-dimensional environment which summarizes the raw data is the SVM's central concern.

3.2 K Nearest Neighbors Classifier

K Nearest Neighbors (KNN) Classifier is the most widespread application and it can be used for both classification and regression. It is a flexible method that will be accustomed to impute missing values and resample datasets.

3.3 Decision Tree Classifier

A Decision Tree is used to follow a set of rules and make decisions. It can perform both classification and regression. It uses quite a few strategies to determine whether or not to break up a node into or greater sub-nodes. In different words, we are able to assert that the purity of the node will increase with recognition of the favored variable.

3.4 Random Forest Classifier

Random forest consists of a large number of individual decision trees that operate as an ensemble. It makes use of quite a few samples to generate choice trees, the use of the bulk of them for categorization and the common of them for regression

3.5 Logistic Regression

Logistic Regression is a statistical analytic approach that expects a binary conclusion, which includes sure or no, zero or 1, proper or false, etc., primarily based totally on earlier observations of a statistics collection.

3.6 XGBoost

XGBoost is a famous and a hit open-supply model of the gradient boosted timber technique (eXtreme Gradient Boosting). In order to expect a goal variable, the supervised mastering approach called gradient boosting combines an ensemble of estimates from numerous weaker and less difficult models.

3.7 Voting Classifier

A vote casting classifier is a selected sort of gadget getting to know an estimator that builds some of base fashions or estimators after which generates predictions via means of averaging their output. In addition to the aggregating criterion, every estimator output may also acquire a vote.

3.8 AdaBoost Classifier

An AdaBoost classifier, a form of meta-estimator, starts by means of becoming a classifier to the preliminary dataset. The equal dataset is then used to suit extra copies of the classifier

with the weights of times that have been mistakenly categorized changed in order that next classifiers could be aware of tough cases

3.9 Bagging Classifier

A bagging classifier is an ensemble meta-estimator that applies primary classifiers separately to arbitrary subsets of the unique dataset, then combines the person forecasts to get a very last prediction.

3.10 Naive Bayes

The Naive Bayes Classifier, one of the best and simplest type algorithms, helps the introduction of fast gadget mastering fashions which are able to make specific predictions. It is primarily based totally on opportunity fashions that make giant independence assumptions. Often, the independence assumptions don't have any effect on reality.

4. DATASETS

4.1 Health *Care - Heart Disease*

One of the illnesses that can be harmful is heart disease, which has recently received a lot of attention in clinical research. The difficult chore of diagnosing a heart attack can provide automated predictions about the patient's cardiac status to increase the efficacy of later treatment. The physical examination, signs, and symptoms of the patient are frequently used to make the diagnosis of heart disease. Numerous factors, such as smoking, body cholesterol, family history of the disease, obesity, high blood pressure, and inactivity, affect the chance of developing heart disease.

4.2 Agriculture - Crop Yield Production

One of agriculture's most difficult problems is predicting crop yields. It is essential for international, regional, and local policymaking. Crop output is anticipated based on agricultural, soil, climatic, and environmental factors, among others. Crop prediction often employs decision - making algorithms to extract critical crop characteristics. Precision agriculture provides the benefit of boosting agricultural output and quality while having a lower environmental effect.

4.3 Education - Student Placement Performance

placements are very important in student career setting a placement prediction gadget may be used to evaluate a pupils suitability for a given profession role because there are such a lot of pupil facts on the college its miles exceptionally difficult to manually expect precise traits we solved the difficulty with the guide method the use of ml algorithms by taking into consideration several traits which include cgpa technical skills coding skills and conversation skills the location prediction approach aids withinside the green filtering of students predict the chance of pupil placement overall performance the use of diverse machine learning ml techniques.

5. METHODOLOGY

In the study three different datasets are used. First dataset used is heart disease with 304 rows and 14 columns. The Second one is crop yield Production with 246092 rows and 7 columns. The third one is student placement performance with 215 rows and 15 columns.

The analysis has been performed in the jupyter notebook platform running on with 3GB RAM installed. As input for the feature extractor and classification algorithm, three data sets are used. The datasets run through several pre-processing blocks in succession.

Figure below shows how these datasets were classified using this methodology.

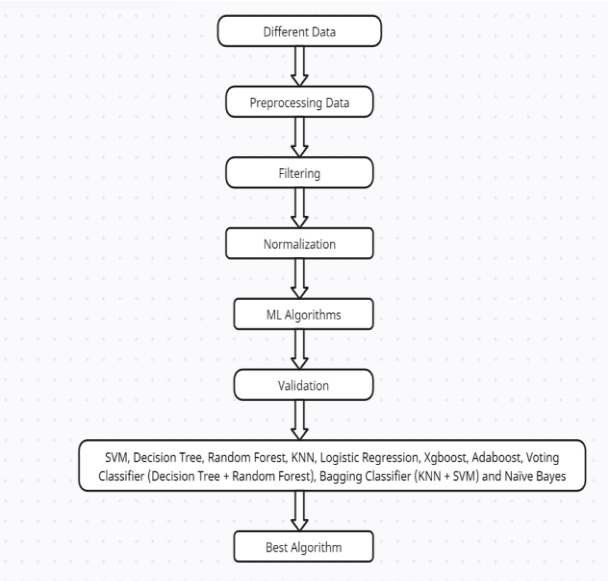

Fig. 1 Proposed System Flow

In Jupyter Notebook, we must first import the libraries required such as pandas, NumPy, matplotlib, etc., then import datasets into the respective jupyter notebook. After importing the dataset, it undergoes pre-processing techniques. Before we should find the missing data, now encode categorical data. After filling in the missing data and converting the data, we should split the entire dataset into both training and testing sets and do feature scaling.

6. RESULTS AND DISCUSSION

We performed training and testing of all 10 different supervised learning algorithms on different datasets, the following are the results obtained.

In this study three different datasets were used. The 10-cross validation strategy is used to approve and evaluate classifier execution. The classifier models with the highest accuracy are chosen for classification after computing the accuracy of each fold. All the accuracy of the three data sets is presented in the table. According to the table the ten algorithms applied on the data sets were applied on 80% train and 20% test. We got different accuracy for various algorithms on three datasets.

Heart Disease:

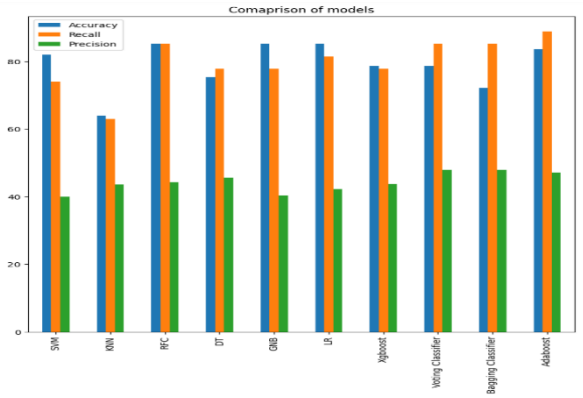

Fig. 2 Outputs of Heart Disease Dataset

From the above figure, we have taken the heart disease dataset in the domain of Healthcare. For this dataset we had applied all the ten algorithms of supervised learning and we got the accuracies, recall, precision respectively. From this we can observe that the Logistic Regression algorithm gives the highest accuracy.

Crop Yield Production:

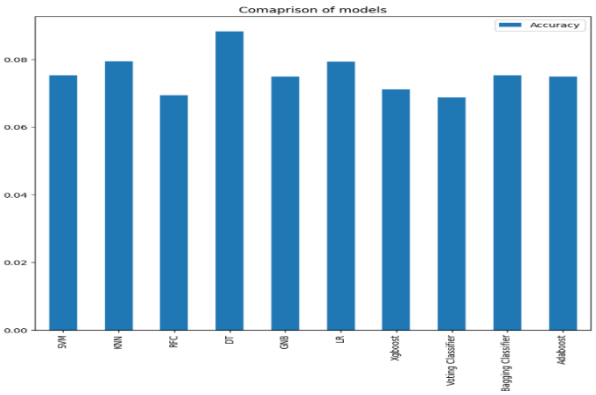

Fig. 3 Outputs of Crop Yield Production Dataset

From the above figure, we have taken the crop yield dataset in the domain of agriculture. For this dataset we had applied all the ten algorithms of supervised learning and we got the mean square errors respectively. From this we can observe that the Adaboost algorithm gives the least mean square error. Which means Adaboost performs well.

Student Placement Performance:

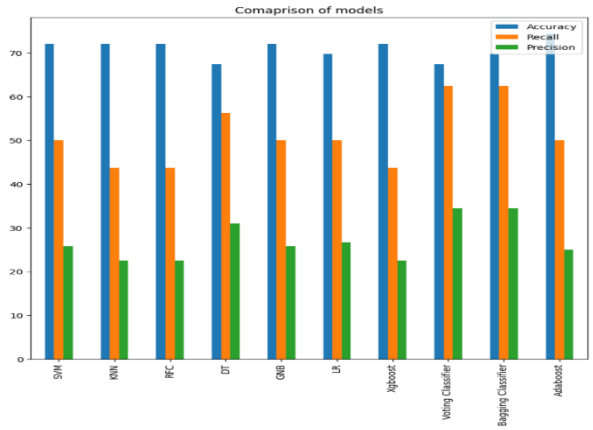

Fig. 4 Outputs of Student Placement Performance Dataset

From the above figure, we have taken the student performance dataset in the domain of education. For this dataset we had applied all the ten algorithms of supervised learning and we got the accuracies, recall, precision respectively. From this we can observe that the Adaboost and Gradient Naive Bayes algorithms give the highest accuracy.

ALGORITHMS	HEART DISEASE ACCURACY	CROP YIELD MEAN SQUARED ERROR	STUDENT PERFORMANCE ACCURACY
SVM	81.9%	75.3%	72%
KNN	63.9%	79.3%	72%
RFC	81.9%	69.4%	72%
DT	73.7%	88.2%	69.7%
GNB	83.6%	74.8%	74.4%
LR	85.2%	74.9%	69.7%
Xgboost	78.6%	79.3%	72%
Voting Classifier	80.3%	71.1%	72%
Bagging Classifier	72.1%	75.3%	69.7%
ADA boost	83.6%	68.7%	74.4%

Fig. 5 Accuracy Table

The above figure gives us information about three different datasets along with ten different algorithms on each dataset respectively in the form of a table. For the heart disease dataset we have calculated the accuracy percentage. For the crop yield dataset we have calculated the least mean square error percentage. For the student performance dataset we have calculated the accuracy percentage.

For Heart Disease dataset *Logistic Regression*, *Naive Bayes*, *Adaboost*, *Random Forest*, *SVM* gave amazing results with *>80%* of accuracy. For Crop Yield data set *Adaboost*, *Random Forest* produced results with *<70%* of mean squared error. For the Student Performance data set *SVM*, *k-NeighborsClassifier*, *Random Forest*, *Naive Bayes*, *Xgboost*, *Voting Classifier*, *Adaboost* gave results *>70%* of accuracy.

7. CONCLUSION

Our study aims to suggest the best algorithm for each domain of datasets by implementing a model which combines the accuracies from various Supervised Machine Learning Algorithms on different domain datasets. This study has made use of ten Machine Learning (ML) algorithms on each of the three different domain datasets to obtain the final analysis of algorithms and achieved the logistic regression algorithm for heart disease dataset, Adaboost algorithm for crop yield dataset and Adaboost algorithm for student performance dataset as best algorithms for each the domains. In our study we also observed and made analysis that Adaboost algorithm gives the overall better performance in any domain chosen.

8. REFERENCES

[1] Sharma, L., Gupta, G. and Jaiswal, V., 2016, December. Classification and development of tools 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] Chanumolu, S.K., Gupta, A., Jaiswal, V., 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.

[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.