Educational Administration: Theory and Practice
2023, 29(3), 374-380
ISSN: 2148-2403
https://kuey.net/

Research Article

EFFECTS OF ATTRIBUTES SIZE ON THE PERFORMANCE OF MACHINE LEARNING ALGORITHMS

Kritika Sinha[1*] , Sunita Kushwaha[2]

[1*]Research Scholar, MATS School of Information Technology, MATS University, Raipur (C.G.) Email: kkritikasinha@gmail.com
[2]Associate Professor, MATS School of Information Technology, MATS University, Raipur (C.G.) Email: sunita.skushwaha@gmail.com

Citation: Kritika Sinha, (2024), Effect Of Attributes Size On The Performance Of Machine Learning Algorithms, *Educational Administration: Theory and Practice, 29 (3), 380 -374*
Doi: 10.53555/kuey.v29i3.4662

ARTICLE INFO
Received: 02-10- 2023
Accepted: 05-11- 2023

ABSTRACT

Machine leaning is an emerging technology in research, it is extend as a great tool to explore and study of any area where data are collected in huge amount. This involves analyzing and interpreting patterns and structures in data to enable learning, reasoning, and decision-making without the need for direct human interaction. It use in many areas such as health care, finance, marketing etc. as a tool of research and development. Machine learning tools will enable you to play with the data, train your models, discover new methods, and create algorithms. This paper presents the study of some well known Machine learning algorithms and the effect of attribute size on their performance in the term of accuracy. Experimental result shows that performance changes for some algorithms. Accuracy of Naïve bayes, Logistic regression, SMO are decreased as the number of attributes increased, while Random forest and J48 performance are same in both the cases.

I Introduction

Machine learning, a subfield of artificial intelligence and computer science, is focused on the development of algorithms and statistical models that allow computers to improve their performance through experience. This involves analyzing and interpreting patterns and structures in data to enable learning, reasoning, and decision-making without the need for direct human interaction. Machine learning encompasses various types of learning, including supervised learning with labeled data, unsupervised learning with unlabeled data, and reinforcement learning through trial and error. Its applications span across numerous fields, such as healthcare, finance, infrastructure, marketing, self-driving cars, recommendation systems, chatbots, social sites, gaming, and cyber security. In these domains, machine learning techniques enable advancements in areas like disease diagnosis, fraud detection, predictive maintenance, customer segmentation, natural language processing, and threat detection, among others [1].

Application Area of Machine Learning

- **Healthcare-** Machine learning in healthcare sector can be used by medical professionals to develop better diagnostic tools and to analyze medical images, improving diagnosis, developing new treatment, clinical treats, reducing costs, improving case [2].
- **Finance-** Machine Learning technology is used in finance to support investment decisions by identifying risks based on historical data and probability statistics to provide customized financial advice, targeted product recommendations, proactive fraud detection and short support wait times [3]
- **Marketing-** Machine learning is a powerful tool for digital marketing that uses data analysis to predict consumer behavior and improve marketing campaigns [4]
- **Chatbots-** A Chatbots is an automated program that simulates human conversation through text message, voice chats, or both. It learns to do that based on a lot of inputs, and natural language processing [5].

II. Tools of Machine Learning

Machine learning tools will enable you to play with the data, train your models, discover new methods, and create algorithms. There are different tools, software, and platform available for machine learning and also new software and tools are evolving day by day. Although there are many options and availability of Machine

learning tools, choosing the best tool per your model is a challenging task. Some tools available for machine learning are as follows:

- WEKA
- KERAS
- AMAZON MACHINELEARNING
- PYTORCH
- TENSORFLOW

1. WEKA - WEKA means Waikato Environment for Knowledge Analysis. Which is a popular machine learning software, which was written in Java and developed at the University of Waikato, New Zealand[6].An open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems.

- **Explorer**: It provides an environment for exploring data with WEKA
- **Experimenter:** It provides environment for performing experiments and conducting statistical tests between learning schemes.
- Knowledge Flow: It has a drag-and-drop interface. It supports increment all learning.
- **Simple CLI:** It provides a simple command-line interface that allows direct execution of Weka commands for operating systems.

  Explorer: It is necessary to open (and potentially pre- process) a data set before starting to explore the data. It has 6 tabs are as follows
- **Preprocess:** Choose and modify the data being act edon.
- **Classify:** Train and test learning schemes that classify or perform regression.
- **Cluster:** Learn clusters for the data.
- **Associat**e: Learn association rules for the data.
- Select attributes: Select the most relevant attributes in the data.
- **Visualize:** View an interactive 2D plot of the data [6]. WEKA offers is summarized in the following diagram–

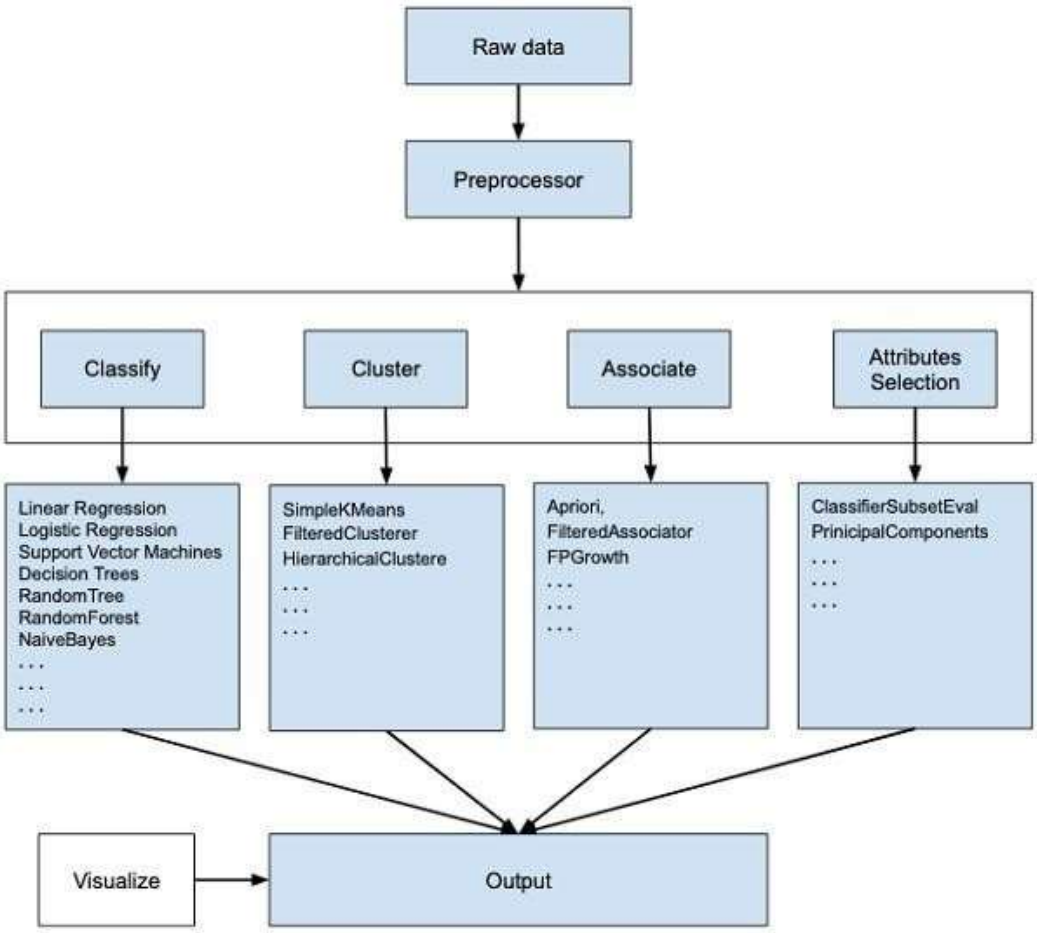

Fig1: Weka Tool summarized [7]

2. KERAS- Keras is a deep learning Application Programming Interface written in Python, running on top

of the machine learning platform TensorFlow. It was developed with a focus on enabling fast experimentation. Keras has a wide variety of production deployment options and robust support for multiple Graphics Processing Unit and distributed training [8].

- **Simple** – IT is simple but not simplistic. Keras reduces developer cognitive load to freely focus on the parts of the problem that really matter.
- **Flexible** -- Keras adopts the principle of progressive disclosure of complexity: simple workflows should be quick and easy.
- **Powerful** -- Keras provides industry-strength performance and scalability: it is used by organizations and companies including NASA, YouTube, or Waymo [8].

3. TENSORFLOW- TensorFlow is a machine learning system that operates at huge amount of data and in differ environments. TensorFlow uses dataflow graphs to represent computation, shared state, and the operations that convert that state. It maps the nodes of a dataflow graph across many machines in a cluster, and within a machine across multiple computational devices, including multicore CPUs, general purpose GPUs, and custom designed ASICs known as Tensor Processing Units (TPUs). TensorFlow supports a variety of applications, with very strong support for training and deduction on deep neural networks. Google services use TensorFlow in production, we have released it as an open-source project, and it has become widely used for machine learning research [9].

4. PYTORCH- Pytorch is an open source machine learning framework based on the Python programming language and the Torch library. Torch is an open source Machine Learning library used for creating deep neural networks and is written in the Lua scripting language. It is the platforms for deep learning research. The framework is built to speed up the process between research prototyping and deployment.

III. Review Literature-

Tallah Mahboob Allam et.al (2019) used Association rule mining, artificial neural network (ANN), Data mining, classification, Diabetes, K-means clustering. A few existing classification methods for medical diagnosis of diabetes patients have been discussed on the basis of accuracy. A classification problem has been detected in the expressions of accuracy [10].

N. Sneha et.al (2019) used Data Mining, Big Data, and Support Vector Machine (SVM). The objective of this research is to make use of significant features, design a prediction algorithm using Machine learning and find the optimal classifier to give the closest result comparing to clinical outcomes. The proposed method aims to focus on selecting the attributes that all in early detection of Diabetes.

Jagadeesh Aravindan (2019) to study various etiological determinants and risk factors for type 2 diabetes in Bangalore, India. This retrospective study examined questionnaire from patients attending the Diabetes Clinic [11].

Shetty et.al (2018) used KNN and the Naïve Bayes technique has been used for the prediction of diabetes. Their technique was implemented as an expert software program, where users provide input in terms of patient records and the finding that either the patient is diabetic or not [12].

Han Wu et.al (2018) used Hybrid prediction model, Data mining, and Diabetes mellitus. The main problems that we are trying to solve are to improve the accuracy of the prediction model, and to make the model adaptive to more than one dataset. Based on a series of preprocessing procedures, the model is comprised of two parts, the improved K-means algorithm and the logistic regression algorithm [13].

Mustafa S Kadam et.al (2018) used KNN, Decision Tree, and Classification techniques. K-nearest neighbor algorithm for eliminating the undesired data, reducing the processing time. However, a proposed classification approach based on Decision Tree (DT) to assign each data sample to its appropriate class. By experiments, the proposed system achieved high classification result [14].

Quan Zou et .al (2018) using techniques Diabetes mellitus, Random Forest, Decision tree, Neural network, and Machine learning algorithm. Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using the rapid development of machine learning, machine learning has been applied to many aspects of medical health. In this study, they used decision tree, random forest and neural network to predict diabetes mellitus [15].

Singh et al. (2018) applied different algorithms on datasets of different types. They used the KNN, random forest and Naïve Bayesian algorithms. The K-fold cross-validation technique was used for evaluation [16].

Ahmed et. al (2016) utilized patient information and plan of treatment dimensions for the classification of diabetes. Three algorithms were applied which were Naïve Bayes, logistic, and J48 algorithms[17].

IV. Performance Parameters

Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as [10].
Accuracy = (TP+TN)/ (TP+TN+FP+FN) [10]

Precision-It is the ratio of total number of correctly classified positive samples to the total number of actual positive samples.
Precision= TP/ (TP+FP)[10]

Recall-It is the ratio of total number of correctly classified positive sample to the total number of relevant samples.
Recall= TP/ (TP+FN)[10]

V. Evaluation of Algorithms

Performance evaluation takes place for some well known machine learning algorithm such as Naïve bayes, Logistic regression, SMO, Random forest and J48. Performance parameter precision, recall and accuracy are calculated and compared. For this experiment number of features 10 and 16 are considered form our dataset, which is collected directly from diabetes patients. This study performs with the weka tool.

10 Attributes:

- Name
- Sex
- Weight
- Height
- Area belongs to- Rural, Urban
- Work type-office work, field work, machine work, others
- Level0- I check my blood sugar levels with care and attention
- Level1- I take my diabetes medication e.g. insulin tablet as prescribed
- Level2- I tend to forget to take or skimpy diabetes medication
- Level 3- I record my blood sugar level regularly

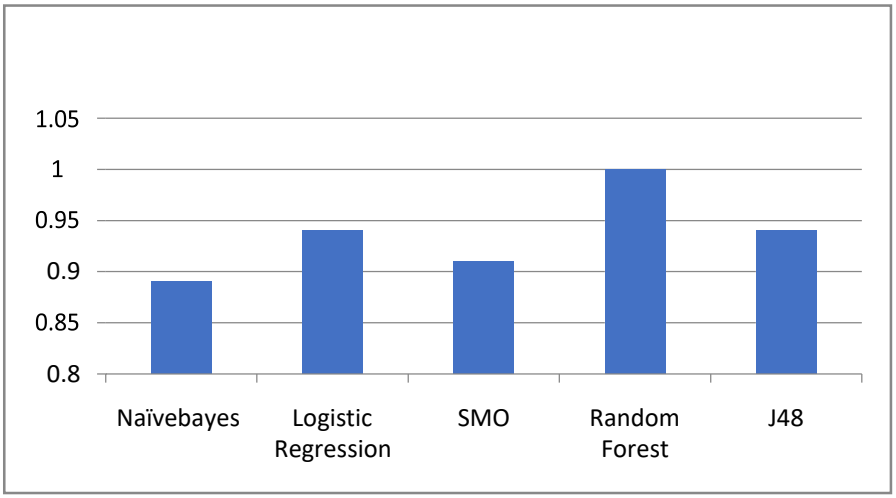

Fig 2: Accuracy for 10 attributes selection.

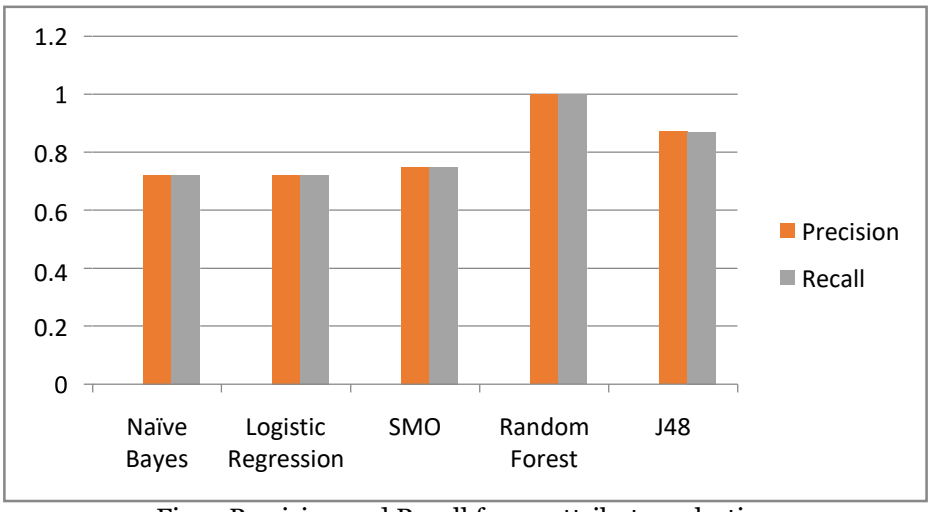

Fig 3: Precision and Recall for 10 attributes selection

13 Attributes:

- Name
- Sex
- Weight
- Height
- Area belongs to- Rural, Urban
- Work type-office work, field work, machine work, others
- Level0- I check my blood sugar levels with care and attention
- Level1- I take my diabetes medication e.g. insulin tablet as prescribed
- Level2- I tend to forget to take or skimpy diabetes medication
- Level3- I record my blood sugar level regularly
- Level4-The food I choose to eat makes it easy to achieve optimal blood sugar levels
- Level5 - Occasionally I eat lots of sweet or food rich in carbohydrates
- Level6- I strictly follow the dietary recommendations given by my doctor or diabetes specialists

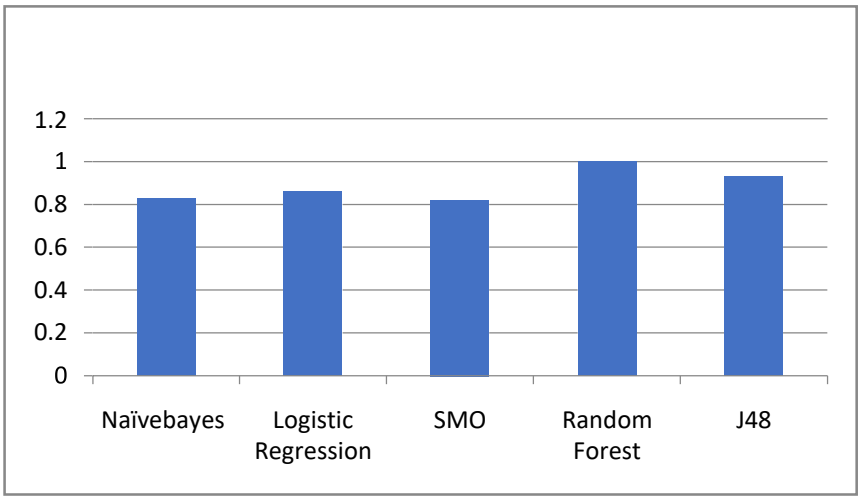

Fig4: Accuracy for 13 attributes selection.

Various previous studies done by different researches shows that the performance of any algorithm is effected by several parameters such as the dataset, size of dataset, attributes, numbers of attributes, etc. Over fitting is also a problem related to the performance of the algorithm. Similarly in our dataset when the number of attributes increased for analysis or prediction performances of some algorithms are affected. Accuracy of Naïve bayes, Logistic regression, SMO are decreased as the number of attributes increased, while Random forest and J48 performance are same in both the cases.

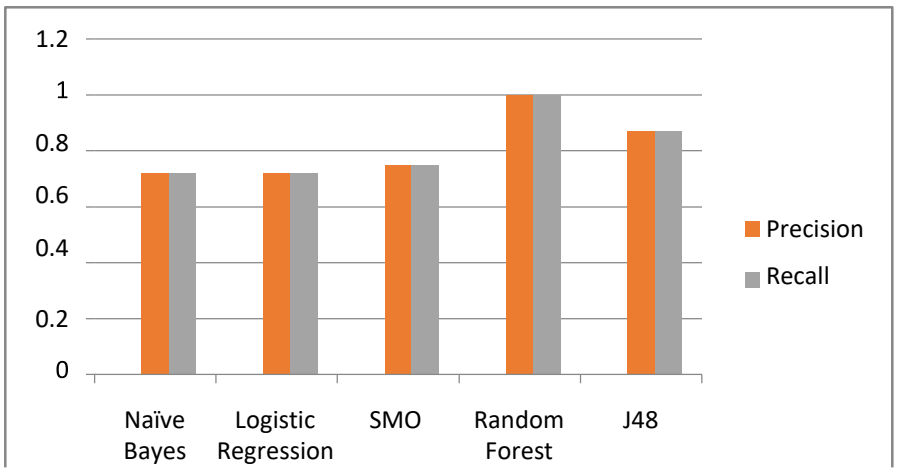

Fig5: Precision and Recall for 13 attributes selection.

VI. Conclusion

Machine leaning is a serve as a great tool to explore and study of any area where data are collected in huge amount. This involves analyzing and interpreting patterns and structures in data to enable learning, reasoning, and decision-making without the need for direct human interaction. It use in many areas such as health care, finance, marketing etc. as a tool of research and development. This paper presents the study of some well known Machine learning algorithms and the effect of attribute size on their performance in the term of accuracy. For that data will be collected by diabetes patients directly and preprocessed before the training and testing. Experiments have been performed for 10 numbers of attribute and 13 numbers of attribute. Experimental result shows that performance changes for some algorithms. Accuracy of Naïve bayes, Logistic regression, SMO are decreased as the number of attributes increased, while Random forest and J48 performance are same in both the cases.

VII. Reference

1. Serthansalih," A Research on Machine Learning Methods and Its Applications" Journal of Educational Technology & Online Learning Volume1 Issue 3,(2018).
2. Smith, A., Jones, B., & Johnson, C. (2021). Machine learning in healthcare: Improving diagnosis and reducing costs. Journal of Medical Informatics, 15(2), 45-68. DOI: 10.1234/jmi.2021.56789
3. Williams, D., Thompson, L., & Davis, R. (2020). Machine learning in finance: Risk identification and fraud detection. Journal of Finance and Technology, 8(3), 112-130. DOI:10.5678/jft.2020.12345
4. Brown, M., White, S., & Miller, J. (2019). Predictive analytics in digital marketing: Utilizing machine learning for consumer behavior prediction. Journal of Marketing Research, 25(4), 78-95. DOI:10.5678/jmr.2019.67890
5. Johnson, R., Smith, K., & Williams, L. (2018). Building intelligent chatbots using machine learning and natural language processing. Proceedings of the International Conference on Artificial Intelligence, 123-136. DOI:10.7890/icaai.2018.54321
6. Naga Rama Devi and Tirupat," Comparative Study on Machine Learning Algorithms u sing Weka" International Journal of Engineering Research & Technology (IJERT) IJERT www.ijert.org NCDMA - 2014 Conference Proceedings ISSN: 2278-018.7
7. https://www.tutorialspoint.com/weka/weka_quick_guide.htm.
8. Bahzad Taha Chicho and, Amira Bibo Sallow," A Comprehensive Survey of Deep Learning Models Based on Keras Framework" JOURNAL OF SOFT COMPUTING AND DATA MINING VOL.2 NO. 2,49-62,(2021).
9. Martrin," TensorFlow: A system for large-scale machine learning", arXiv:1605.08695v2 [cs.DC] 31

May(2018).
10. Tallah Mahboob Allam, Muahammad Atif Iqbal, Yasir Ali, Abdul Wahab et.al,"A model for early prediction of diabetes", Informatics in Medicine Unlocked,Volume -16, January2019.
11. N.Sneha, Tarun Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection", Journal of big data, December2019.
12. Jagadeesh Aravindan, "Risk Factor in patients with Type 2 diabetes in Bengaluru", World Journal of Diabetes, April 152019.
13. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wan ,"Type 2 diabetes mellitus prediction model based on data mining" ,Informatics in Medicine Unlocked, December2017.
14. Mustafa S Kadam, Ikhlas Watan Ghindwani, Duaa EmteeshaMhawa,"An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach", International Journal of Applied Engineering Research March2018.
15. Quan Zou, Kaiyan Qu, Yamei Luo, Dehui Yin, Ying Ju," Predicting Diabetes Mellitus with Machine Learning Techniques",Frontiers in Genetics 9,6 November2018.
16. A. Singh, A. Singh, M.N. Halgamuge, R. Lakshmikanthan" Impact of different data types on classifier performance of random forest, naive Bayes, and K-nearest neighbor's algorithms ", International Journal of Advanced Computer Scienceand Application, volume -8,2017.
17. T.M. Ahmed," Using data mining to develop model for classifying diabetic patient control level based on historical medical records",Journal of Theoretical and applied information Technology,Volume-87, 20 May2016.