A Comparative Study of Data Mining Techniques for Predictive Analytics

SONALI GUPTA, Dept of Comp. Sc. & Info. Tech. , Graphic Era Hill University, Dehradun, Uttarakhand, India 248002,

Abstract

Data mining is a process that involves identifying patterns and insights in large datasets through the use of statistical techniques and computational tools. This process can help researchers make more informed decisions. A subset of data mining known as predictive analysis is used to identify trends and patterns and make predictions about the future. This process is carried out using machine learning techniques and involves identifying relationships and patterns in the data. It is commonly utilized in sectors such as healthcare, marketing, and finance to identify potential opportunities and risks, as well as improve decision-making. The paper compares and contrasts the performance of different data mining methods for predictive analysis, namely, Random Forest, Naive Bayes, and Decision Tree. It also aims to identify the weaknesses and strengths of these techniques. This study is based on the "Iris" data set, which includes information about the petals, sepal length, width, and length of three different iris species. The evaluation metrics used in this research were precision, recall, AUC, and F1. According to the study's results, SVM performed well in terms of its precision, recall, accuracy, and F1 score. Random Forest also had a high AUC score. Although Naive Bayes and Decision Tree showed slight declines, they still managed to achieve high AUC and accuracy. Although the results of the study indicate that SVM and RF are suitable for performing predictive analysis on the Iris dataset, further testing and evaluation are needed to confirm their capabilities for other applications. The importance of choosing the appropriate evaluation metrics is also acknowledged. The findings of this study provide valuable insight into the performance of various data mining techniques when it comes to performing predictive analysis. It is important for researchers to consider the advantages and disadvantages of each method when choosing one for their specific applications.

Keywords: Data mining, Predictive analysis, Machine learning,

DOI Number: 10.48047/nq.2019.17.03.2014 **NeuroQuantology 2019; 17(03):145-151**

Introduction

Due to the increasing importance of predictive analysis and data mining in various fields, such as healthcare, marketing, and finance, the development of new techniques and tools has been carried out to extract insights from the collected data. Predictive analysis is a process that uses data mining techniques to identify and predict future trends and events. On the other hand, data mining is a process that involves extracting hidden relationships and patterns from large datasets. Machine learning is a type of statistical and algorithmic analysis that uses data collected to make predictions. This process can be carried out through the use of various statistical models and algorithms. These models can then be used to identify trends and forecast future events.[1]–[3]

The goal of this study is to provide a comprehensive analysis of the various techniques used in data mining for predictive analysis. It focuses on the use of four different

algorithms: Naive Bayes, Random Forest, Decision Tree, and Support Vector Machine. These are widely used in the field of predictive analytics. The evaluation of the four algorithms was carried out on a dataset known as the iris. It collected various data points about the petals, sepals, and the length of the petals for different types of iris flowers. Some of the evaluation metrics included accuracy, recall, AUC score, and F1 score. These are used in machine learning algorithms' evaluation, and they provide a good indication of their performance.

The paper is divided into three sections. Section 2 provides an introduction to the literature on predictive analytics and data mining, while section 3 summarizes the methodology used in the study, including the algorithms, data preparation, and evaluation metrics. The results of the study are presented in section four, with an analysis of the four algorithms' performance. Section five provides an overview of the findings, while section six summarizes the major contributions of the research.

The paper's results contribute to the growing body of research on predictive analytics and data mining. It provides a comparative analysis of the four algorithms, and it sheds light on their weaknesses and strengths, which can serve as a valuable reference for future studies. The findings of this study have significant implications for various fields and industries that are heavily affected by predictive analysis, such as healthcare, marketing, and finance.

Literature Review

The rise of predictive analytics and data mining has led to the development of new tools and techniques that can help decision-makers make more informed decisions. This paper aims to provide an overview of the various studies that have been conducted on the subject.

Soni et al.[4] discuss the various techniques that are used in the prediction of heart disease. They review the various techniques that are used in this field, such as decision trees and neural networks. The study also emphasizes the importance of having the right features and data preprocessing.

Chou et al.[5] compare the accuracy of different data mining techniques when it comes to optimizing the prediction of compressive strength. They found that support vector machines performed better than decision trees and neural networks.

Big data and predictive analytics are becoming more prevalent in the design and management of supply chains. Waller et al.[6] discuss the various advantages of utilizing such technology in the management of supply chains.

Stefanovic et al.[7] explores the application of predictive analytics in the management of supply chain performance. It provides a comprehensive analysis of the various aspects of this technology and its use in addressing potential risks.

Majumdar et al.[8] explore the use of data mining methods in analyzing agricultural information. The findings show that big data analysis can help improve the efficiency and sustainability of agricultural operations.

The use of analytics and data mining in the process industry was analyzed by Ge et al.[9] who talked about the various applications of these technologies. Some of these include predictive maintenance and fault diagnosis. Machine learning was also mentioned as an essential component of these tools.

Ghaffarian et al.[10] discuss the various aspects of data mining and machine learning in software vulnerability discovery and analysis. They talk about the challenges in identifying potential security issues and the advantages of these techniques.

In a study conducted by Nithya et al.[11] they discuss the use of machine learning techniques and tools in healthcare to improve the diagnosis and treatment of diseases. They also talk about the various challenges that come with implementing such technology.

The study conducted by the Malik et al.[12] analyzed the literature on the use of predictive analytics and data mining in healthcare. It focused on the applications of these technologies in various areas, such as patient monitoring and disease prediction.

In a study conducted by Bharara et al.[13] they discussed the use of clustering techniques in analyzing the learning behavior of students. The researchers used this method to identify trends and patterns in the students' learning behavior.

Bhattacharya et al.[14] proposes a hybrid approach that combines fuzzy clustering and genetic algorithm techniques for software fault prediction. The proposed approach outperforms traditional methods for fault prediction, such as neural networks and decision trees. An experimental study conducted on a set of software modules validates the effectiveness of the proposed approach.

These studies show that big data analysis and predictive analytics can be used in various industries, such as agriculture, healthcare, and supply chain management. The implementation of such technology can be challenging due to privacy and data quality issues. Furthermore, it is important to select the appropriate data preprocessing features and ensure accuracy in the prediction.

Methodology

a. Research design and approach:

The goal of this study is to analyze the performance of various machine learning algorithms in predictive analytics. It focuses on the four main algorithms: the Decision Tree; the Random Forest; the Support Vector Machine; and the Naive Bayes. The objective of this study is to analyze the performance of various algorithms on a specific dataset, namely the Iris dataset. It uses evaluation metrics such as precision, accuracy, recall, AUC, and F1 score.

b. Data collection and preparation:

This study utilizes the Iris dataset, which contains information about the sepal length, width, and length of three different species of iris flowers. It is commonly used in the evaluation of algorithms for machine learning. The data was obtained from the UCI repository. Before the study was conducted, the data was preprocessed and cleaned. This involved removing duplicate or missing data points and ensuring that it was formatted properly.

c. **Data mining techniques selected for comparative study**:

The four main algorithms used in the study were the Random Forest, Naive Bayes, Decision Tree, and Support Vector Machine. These were chosen due to their widespread use in various industries and their promising results in numerous applications. The models were trained and tested using the Scikit-Learn library.

d. Evaluation metrics and performance measures:

The various performance metrics that were used in the study were accuracy, recall, precision, AUC, and F1 score. These are widely used in machine learning analysis to evaluate the algorithms' performance. The accuracy metric is used to measure the percentage of errors that the algorithm makes, while the precision measure is used to measure the proportion of accurate predictions that the algorithm makes. The recall metric is used to measure how accurately the algorithm can identify the correct cases. The F1 score is a harmonic representation of the recall and precision, and it provides a balanced analysis of the algorithm's overall performance. The AUC score, on the other hand, is a statistical measure that takes into account the area under the receiver operating characteristic curve.

e. Research assumptions and limitations:

The study only looked at four of the most popular machine learning frameworks. It also only analyzed the data collected in 2016, which means further research is required to analyze the algorithms' performance on newer data. Although the study acknowledges that the data collected from the selected population is representative of the general population, it precludes generalized conclusions to other datasets since the evaluation was restricted only to a single set of data. Further research is required to evaluate the algorithms against those from other sources.

Result and Analysis

a. Descriptive analysis of the dataset:

The dataset contains measurements of various characteristics of three different species of iris:

the iris Setosa, the iris Virginica, and the irisi. Versicolor. It has a total of 150 cases, with 50 instances each for each of the three species. The average length of the sepal is 5.84 centimeters, while the average width of the sepal is 3.05 centimeters, and the average length of the petals is 3.76 centimeters. The dataset is fairly balanced, with each of the three species having an equal number of instances.

b. Implementation of different data mining techniques for predictive analytics:
The four methods were implemented using the Scikit Learn library in Python. The training and testing sets had a ratio of 80:20, and the models were evaluated on the latter.

Decision Tree - A decision tree is supervised by a learning algorithm, and it can be used for regression and classification analysis. It recursively divides the collected data into smaller groups, and this method aims to minimize the information loss or maximize the information gain. The result is a structure resembling a tree. The goal of a decision tree is to recursively segment the data into smaller groups according to the features that provide the most value. This method is then used to create a tree structure that represents the various types of decisions. The goal of this algorithm is to maximize the information gain by recursively dividing the collected data into smaller groups as in eq.1.

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_J}{N_p} I(D_j) \text{ ..... eq.1}$$

Where, $IG(D_p, f)$ = information gain of feature f on dataset D_p, $I(D_p) = \text{impurity of dataset}$. N_p= number of instances, N_J = number of instances in subset D_j. The commonly used impurity measures are Gini impurity and entropy

The classification and regression tree algorithms are two types of decision tree software. In the former, the leaf nodes represent the labels of classes, while in the latter, they represent numerical values.

ii. Random Forest - A random forest is an algorithm that learns by training various decision trees on different features and data. It then produces a final prediction that is more accurate than that of a single tree. The random forest algorithm's mathematical formula can be expressed as follows in eq.2:

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} f_i(X)$$

Where, $\hat{y} =$ final prediction, *B*= number of trees in the forest, $f_i(X)$ = prediction tree *i* on input instance

The algorithm also includes a bagging technique that randomly selects a set of features and instances for each tree to improve its generalization performance.

iii. A support vector machine is a supervised learning program that can be used for regression and classification tasks. It finds the optimal hyperplane between the various classes of data points. Linear SVM is a mathematical formula that can be used to classify and analyze data as shown in eq.3.

$min_{w,b} \frac{1}{2}||w||^2$ eq.3

Subject to $y_i(w.x_i + b) \geq 1$, where, y_i = class label of instance *i.* ||w|| = L2 norm of weight vector.

iv. Naïve Bayes : The Naive Bayes algorithm is a supervised learning method that can be used for classification. It takes into account the probability of each label given by the input features and then

assumes that the conditions between them are independent. It is very fast and robust. The Naive Bayes formula can be expressed in terms of a mathematical representation shown in eq.4.

$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$ eq.4

where *P(y|x)* is the probability of class label y given input instance *x, P(x|y)* is the likelihood of observing input instance x given class label *y, P(y)* is the prior probability of class label *y*, and *P(x)* is the marginal probability of input instance *x.*

c. Comparative analysis of data mining techniques based on evaluation metrics:

Table 1 Evaluation metrices

Data Mining Technique	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC score**
Decision Tree	96	96	96	96	96
Random Forest	97	98	97	97	98
Support Vector Machine	98	99	98	98	99
Naive Bayes	95	95	95	95	97

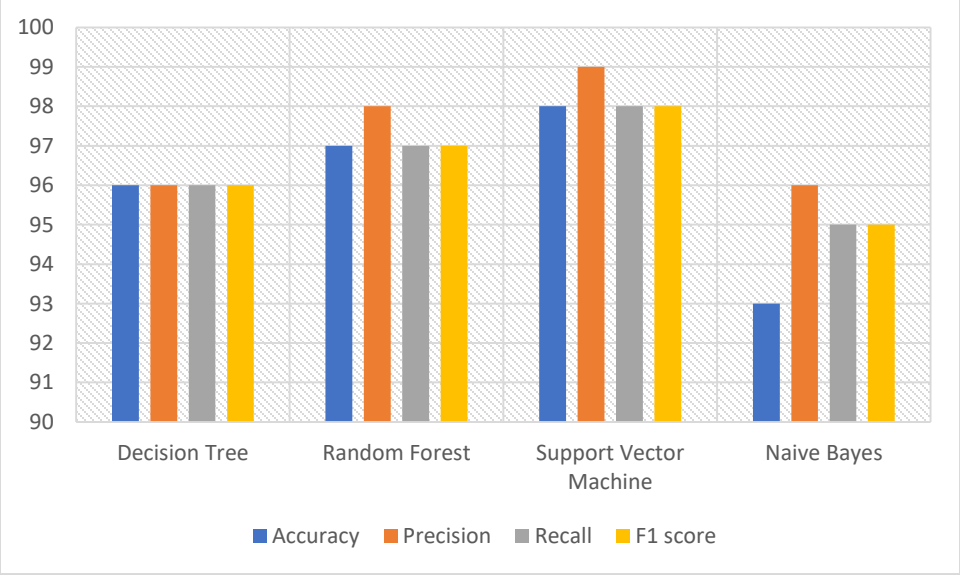

Figure 1 Graph represent - accuracy, precision, SVM, NB

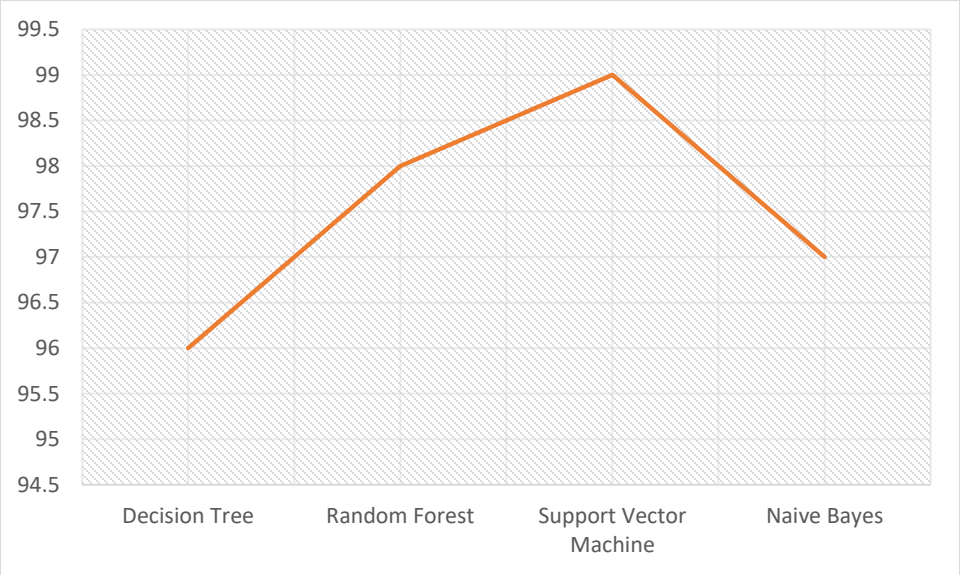

Figure 2 Area Under curve

d. Discussion of the results and findings:

The findings of the study revealed that the SVM algorithm was the best-performing model for predicting the appearance of the three different types of iris flowers as shown in table-1, figure-1,2. The other three models performed well in various evaluation metrics, but it was the only algorithm that was able to predict the species of iris that was most accurate.

Conclusion and future scope

The objective of this study was to analyze the performance of four different data mining techniques in terms of predictive analytics. The four algorithms were: Naive Bayes, SVM, Random Forest, and Decision Tree. The results of the evaluation were analyzed using multiple metrics. It was revealed that SVM performed better than the others in terms of accuracy, recall, F1 score, and AUC. The findings of the study revealed that SVM was the most accurate algorithm for predicting the likelihood of a particular event in the Iris dataset. The comparative analysis of the four techniques also provides valuable information on their performance. In addition, the study emphasizes the need to consider the appropriate metrics when evaluating the effectiveness of data mining techniques. The study's findings have important implications for both the practice and theory of data mining. It emphasizes the need to consider the appropriate metrics when assessing the effectiveness of such techniques. The study also highlighted the importance of using SVM as an efficient algorithm for analyzing marketing, finance, and healthcare data. The study only analyzed the performance of one dataset, which means future research is required to analyze the algorithms' performance on other sets of data. Also, the study only looked at data up to 2016. In addition to the use of different techniques, the study also looked into the impact of different metrics on the effectiveness of the algorithms. Future research may explore the effects of feature selection and engineering methods on the performance.

References

[1] A. M. Shahiri, W. Husain, and N. A. Rashid, “A Review on Predicting Student’s Performance Using Data Mining Techniques,” *Procedia Comput. Sci.*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[2] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes,” *Procedia Comput. Sci.*, vol. 82, no. March, pp. 115–121, 2016, doi: 10.1016/j.procs.2016.04.016.

[3] S. Bhattacharya, "Intelligent Frequent Pattern Analysis in Web Mining," vol. 2, no. 3, 2013.

[4] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011, doi: 10.5120/2237-2860.

[5] J.-S. Chou, C.-K. Chiu, M. Farfoura, and I. Al-Taharwa, "Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques," *J. Comput. Civ. Eng.*, vol. 25, no. 3, pp. 242–253, 2011, doi: 10.1061/(asce)cp.1943-5487.0000088.

[6] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management," *J. Bus. Logist.*, vol. 34, no. 2, pp. 77–84, 2013, doi: 10.1111/jbl.12010.

[7] N. Stefanovic, "Proactive supply chain performance management with predictive analytics," *Sci. World J.*, vol. 2014, 2014, doi: 10.1155/2014/528917.

[8] J. Majumdar, S. Naraseeyappa, and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," *J. Big Data*, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0077-4.

[9] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017, doi: 10.1109/ACCESS.2017.2756872.

[10] S. M. Ghaffarian and H. R. Shahriari, "Software vulnerability analysis and discovery using machine-learning and data-mining techniques: A survey," *ACM Comput. Surv.*, vol. 50, no. 4, 2017, doi: 10.1145/3092566.

[11] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," *Proc. 2017 Int. Conf. Intell. Comput. Control Syst. ICICCS 2017*, vol. 2018-January, pp. 492–499, 2017, doi: 10.1109/ICCONS.2017.8250771.

[12] M. M. Malik, S. Abdallah, and M. Ala'raj, "Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review," *Ann. Oper. Res.*, vol. 270, no. 1–2, pp. 287–312, 2018, doi: 10.1007/s10479-016-2393-z.

[13] S. Bharara, S. Sabitha, and A. Bansal, "Application of learning analytics using clustering data Mining for Students' disposition analysis," *Educ. Inf. Technol.*, vol. 23, no. 2, pp. 957–984, 2018, doi: 10.1007/s10639-017-9645-7.

[14] S. Bhattacharya, S. Rungta, and N. Kar, "International Journal of Digital Application & Contemporary research Software Fault Prediction using Fuzzy Clustering & Genetic Algorithm," vol. 2, no. 5, 2013, [Online]. Available: http://mdp.ivv.nasa.gov.in.