SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS

Prateek Grewal[1], Prateek Sharma[2], Dr Anu Rathee[3], Dr Shikha Gupta[4]
[1,2] *Research Scholar, Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India*
[3,4] *Assistant Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India*

ABSTRACT
Data in its raw form might not mean much but after processing the data and making it more uniform it might reveal a lot of information. By using different types of machine learning algorithms, we can draw a lot of insights. This practice is already being carried out on a very large scale in today's world but as the field of Machine Learning and Artificial Intelligence has advanced a lot, we have so many different algorithms at our disposal but the problem is data can be of many different types and there is no one algorithm that fits the best in every case. Using a complex model might not be useful for a simple dataset or vice versa and this practice might cost a company a lot of time, money and even after that the results might not be the best. Our goal is to depict this and identify which type of Algorithm gives the highest accuracy for which type of Dataset and identify the key factors that influence these algorithms, to demonstrate this we are using IRIS dataset and Wine quality dataset. Based on our research, we conclude that for simple and evenly distributed datasets such as Iris dataset, algorithms like KNN give the best results (95.5% accuracy). For non-uniform simple datasets such as Wine Quality dataset, algorithms like Decision Tree give 100% accuracy and KNN gives the lowest, 82.29%.

KEYWORDS: *Machine Learning, Comparative Analysis, Iris Dataset, Algorithms, KNN, Logistic Regression, Decision Tree, SVM, Naive Bayes, Random Forest, Wine Quality Dataset*

INTRODUCTION

Machine Learning is one of the fastest growing fields and is now used by most of the top companies to help make them better and more informed choices on the basis of the data that, they collect from their customers. Since it is used in so many different fields such as education, medicine, robotics, gaming, etc. The Data that is collected tends to be disparate and can't be studied or used in a similar fashion.

This Study aims to help in classification of Machine Learning Algorithms and provide some insight as to which type of Algorithm would best suit our need depending on the Data that we have. This type of classification can possibly lead to saving a lot of time, money and computational power.

In the current study, we have used commonly known datasets i.e., the Iris dataset and Wine Quality dataset to draw some results and as a proof of concept. Both the datasets are classified into three different classes and have a fairly centered distribution of data (i.e., number of outliers are very less). But the distinction between the two datasets is the uniformity of distribution of data. In the Iris dataset, there is a fairly balanced distribution across all three classes whereas in the Wine Quality dataset, most of the records belong to a particular class and there are very few data points which belong to the remaining two classes. After plotting the datapoints for the two datasets, it was clearly visible that for the Iris dataset, values are well scattered across different classes whereas in the Wine Quality dataset, we observed that most datapoints were overlapping. In case of the Iris dataset, correlation between parameters was higher when compared to the wine quality dataset as shown through their heat maps.

Initially we studied the data by plotting different types of graphs. These plots revealed the similarities and differences between the two datasets as mentioned above. Then we applied six different machine learning classification algorithms i.e., K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, SVM, Random Forest and Naive Bayes. These algorithms are suited for different types of datasets. Our research helped us to identify the models which worked well on evenly distributed uniform data and those which worked well on non-uniform datasets. The approach also helped us to distinguish between the performance of these algorithms using the evaluation metrics of accuracy, precision, recall and F1 score.

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

For the Iris dataset, which is an evenly distributed uniform dataset, KNN gave an accuracy of 95.5% which was the highest amongst all algorithms. For the Wine Quality dataset, which is a non-uniform dataset, Decision Tree yielded a perfect accuracy of 100% and KNN gave the lowest, which was 82.29%.

LITERATURE REVIEW

Kannapiran, T. Et al [1] showed that while using a 75% and 25% split in test and train data on iris dataset, KNN gave 97.5% training accuracy and 96.6% testing accuracy. Prathima,p. Et al [2] tested accuracies of KNN, Decision Tree, SVM and Random Forest on Iris dataset while varying the test and train split between 70-30, 75-25 and 80-20. In all cases SVM performed the best and KNN performed the second best while Decision Tree has the worst performance. Yuanyuan Wu. Et al [3] In iris dataset simply using random forest might not get the best results but with certain modifications it has the potential to outperform all other models used, here random forest was enhanced to create a new GraftedTrees models which outperformed other models such as KNN.

Alghobiri M. et al [4] on a model such as iris dataset it has been found that even advanced Models tend to show a very close results and simplistic models have a higher chance of performing better. To demonstrate such an effect, we have decided to include KNN and a Decision Tree to compare it against random forest. Muhamedyev, R. et al [5] Training of KNN deteriorated when the complexity of dataset increases and the test date is not well separated.

Gupta Y. et al [6] Wine quality dataset has a very high variability amongst the values of parameters we use to predict its quality which has a very high effect on the performance of certain algorithms. Out of all the models used in the study it was found that SVM performed the best for this dataset after greatly reducing the variability of values and removing certain parameters from consideration.

Gupta, M. et al [7] reported that on comparing accuracies KNN, SVM and random forest on Wine Quality dataset, it was found that Random Forest had the highest accuracy and the lowest miscalculation rate. Sharma, N. et al [8] using white wine quality dataset which has similar parameters but different values from ours, it was found that among KNN, Logistic regression, SVM, Random Forest and decision tree models the highest accuracy was attained by Random Forest where decision tree was a close second and the worst accuracy was notes by KNN.

DATA USED

The datasets used for this study were Iris dataset and Wine Quality datasets. A detailed description of the datasets is tabulated and described as below:

Iris dataset

Class	No of samples
Iris Setosa	50
Iris Versicolor	50
Iris Virginica	50

Table 1: Classes and their sample distribution for Iris dataset

The dataset has 6 features, namely – Id, SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, and Species.

Wine Quality dataset

Class	No of samples
Grade A	217
Grade B	1319
Grade C	63

Table 2: Classes and their sample distribution for Wine Quality dataset

The dataset has 13 features, namely – fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality (score between 0 and 10) and grade.

There are several similarities and differences in both the datasets.

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

The similarity between the datasets is that both the datasets are classified into three different classes and have a fairly centered distribution of data. By this we mean that the number of outliers in the data are very less. This saves pre-processing steps as the data does not need to be normalized and cleaned to deal with the outliers.

The distinction between the two datasets is the uniformity of distribution of data.

In the Iris dataset, there is a fairly balanced distribution across all three classes. There are an equal number of data-points for each class in the dataset. The fractional percentage of each class for the Iris dataset is depicted in figure 1 below.

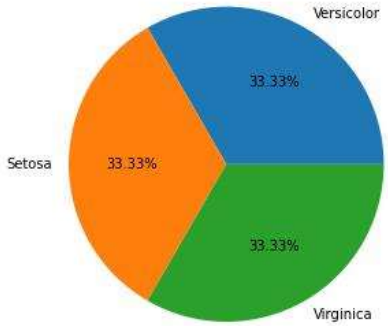


Figure 1: Distribution of data across different classes for Iris dataset

In the Wine Quality dataset, most of the records belong to a particular class - Grade B, and there are very few data points which belong to the remaining two classes- Grade A and Grade C. The data point distribution for each grade for the wine quality dataset is depicted in figure 2 below.

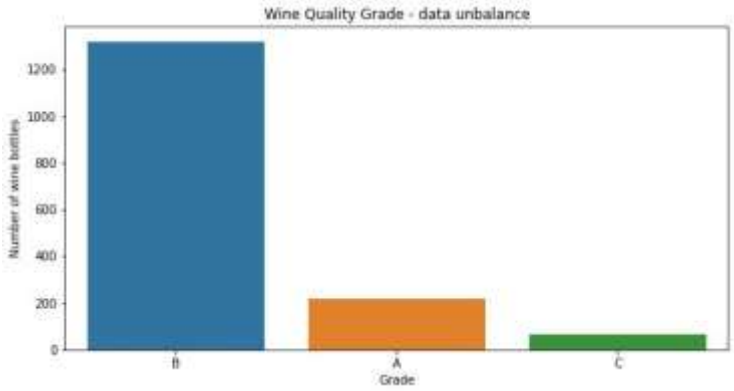


Figure 2: Distribution of data across different classes for Wine Quality dataset

After plotting the datapoints for the two datasets, it was clearly visible that for the Iris dataset, values are well scattered across different classes whereas in the Wine Quality dataset, we observed that most datapoints were overlapping. In case of the Iris dataset, correlation between parameters was higher when compared to the wine quality dataset. The same can be derived from figure 3 below.

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

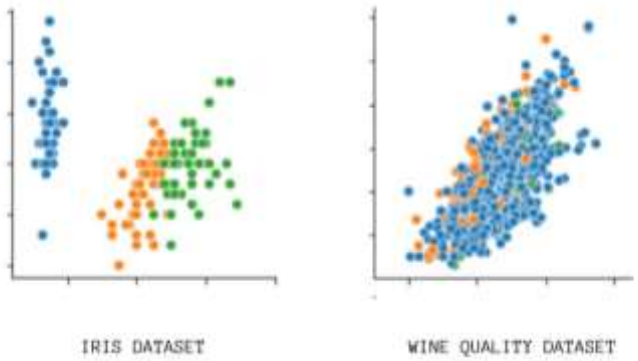


Figure 3: Data point distribution across classes for Iris and Wine Quality dataset

Different Classification Models used in this Study

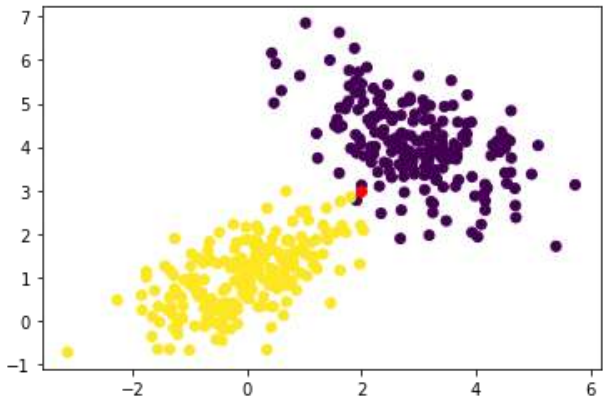
K-Nearest Neighbors: This is a simple machine learning algorithm which comes under the category of supervised machine learning algorithms and it can be used in classification as well as regression problems. The basic assumption in KNN is that similar data points are in close vicinity to each other. We start by calculating the distance of the query point from all the other points in the labelled data. Then we sort these distances in ascending order, the closest to the farthest distance. The 'K 'value in K-Nearest Neighbors is the number of close neighbors we want the model to consider. The next step is to select the K nearest neighbors on the basis of the distances. In regression, our final result is the mean of the corresponding labels of these K nearest neighbors. In classification, the final result is the mode of the labels of the K nearest neighbors.

The advantages of using K-Nearest Neighbors algorithm are:

- Easy to implement
- The model can be used for both- classification and regression problems
- No complexity in making the model

The disadvantages of using K-Nearest Neighbors algorithm are:

- The model becomes slow with increasing data
- Other models are faster and predict better results with large sets of data



Logistic Regression: Logistic regression is a classification algorithm which is used to predict the outcome when the target variable is categorical in nature. It comes under the category of supervised machine learning algorithms. It identifies a decision boundary or a hyper plane between the different categories present in the categorical data. A linear equation combining the input values and coefficients is used to predict the result. The coefficients are determined using the maximum likelihood estimation.

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

Logistic regression is based on the logistic function or the sigmoid function.

$$f(x) = 1/1$$

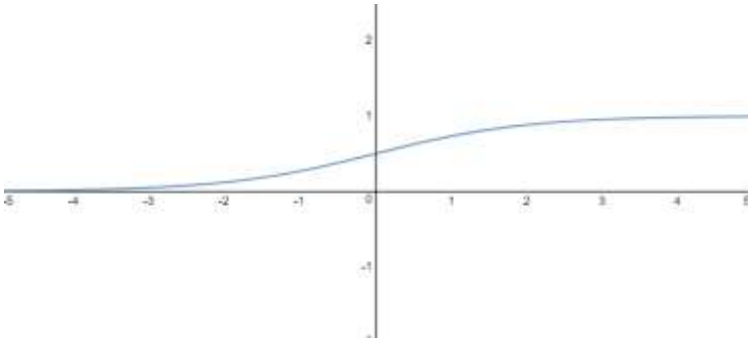


Figure 5: Logistic/Sigmoid Function

The sigmoid function maps any real valued number in {-∞, ∞} to the range {0,1}.
The logistic regression model predicts the probability of the default class for a binary classification. For example, if the probability predicted using logistic regression for the default class is 0.1, it is inferred that the probability for the second class would be 0.9 and hence the query point can be classified to be of the second class.

The advantages of using Logistic Regression algorithm are:

- It makes efficient predictions on linearly separable dataset
- It can be used for data with multiple categories.

The disadvantages of using Logistic Regression algorithm are:

- It can only be used for prediction of discrete variables
- It is prone to overfitting if the dataset is small

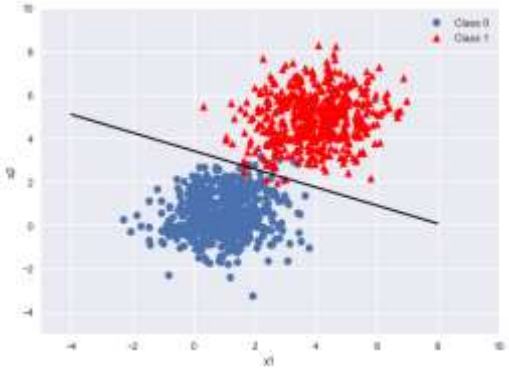


Figure 6: Logistic Regression and its hyperplane

Decision Tree Classifier: The decision tree classifier organizes the data into a simple tree like structure in which the model makes a decision at every node. This model is versatile for simple tasks and it is very popular because we can use decision trees to show how the decision process works.
The algorithm selects the most optimal attribute using Attribute Selection Measures (ASM), makes it a decision node and then splits the dataset around this decision node. This process is recursively carried out for the child nodes until all attributes/ instances have been exhausted.

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

The various Attribute Selection Measures (ASM) are Information Gain, Gini Index and Gain Ratio. Decision tree split using information gain is based on the concept of entropy. Entropy is the measure of randomness of a system. Information gain computes the difference between the entropy before the split and after the split using a particular attribute as the decision node. The goal is to maximize the information gain and hence, the attribute for which the information gain is the maximum is chosen as the decision node.

$$InformationGain$$
$$= H(S) - \sum^{v} \frac{|S_i|}{} H(S_i)$$

Where $H(S) = -\sum P_C log_2 P_C$ denotes Entropy
And P_c = Probability of an arbitrary tuple belonging to class C

The advantages of using Decision Tree Classifier are:

- Mimic human logic and are therefore easy to interpret.
- Efficient in capturing non-linear data patterns
- Due to the non-parametrised approach, there is less preprocessing of the dataset.

The disadvantages of using Decision Tree Classifier are:

- Prone to overfitting
- Slight variations in data can lead to completely different results
- It becomes too complex where there are many attributes involved.

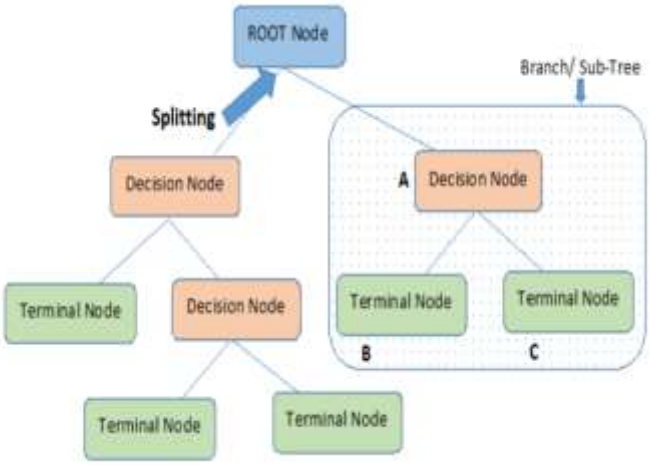


Figure 7: Decision tree classifier

SVM: SVM stands for support vector machine, it belongs to the class of supervised machine learning algorithms. It is commonly used for classification and regression problems. SVM works by trying to find a hyperplane in a N-dimensional space so that we can classify the given datapoints distinctly. The number of dimensions for the hyperplane is dictated by the number of features present in the dataset. For ex: if we consider a dataset that in which there are only 2 input features than the hyperplane will turn out to be a line similarly if there was another input feature added making the total count 3 then the hyperplane would change to a 2-Dimensional plane instead of a line.
We can draw infinite hyperplanes between the datapoint but we try to choose the hyperplane in which we can maximise the separation between the classes present.

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

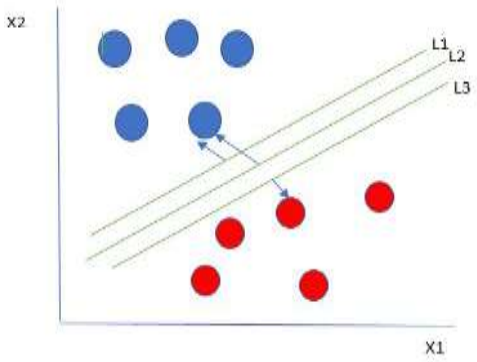


Figure 8: Support Vector Machine (SVM)

Random Forest: It is also a type of supervised ML algorithm that can be used for classification purposed. In this we build multiple decision trees using different sample and then the classification is done using majority votes taken from all the different trees.

One key feature that is highlighted in case of random forest is that it can also be used in case of a dataset that has continuous variables which can be seen in regression problems. In random forest we create several bootstrapped datasets by randomly selecting rows from our original dataset, which then function as independent datasets to create a decision tree for each and then using all of those individual trees we perform the classification.

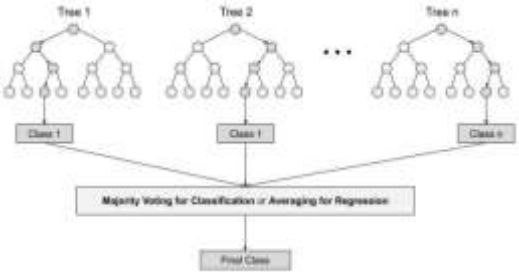


Figure 9: Random Forest classifier

Naive Bayes:

Naive Bayes classifiers is the name given to a family of classification algorithms all of which depend on the Bayes theorem.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Each member of this family shares a common trait in which we make a fundamental assumption in Naive bayes which states that each feature is going to make an independent and equal contribution to out result.

In this we divide the dataset into two parts, the first one is called feature matrix and response vector. The first one i.e. feature matrix contains all the rows of dataset which have the dependent features, The second one i.e response vector contains the prediction/output for each row.

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

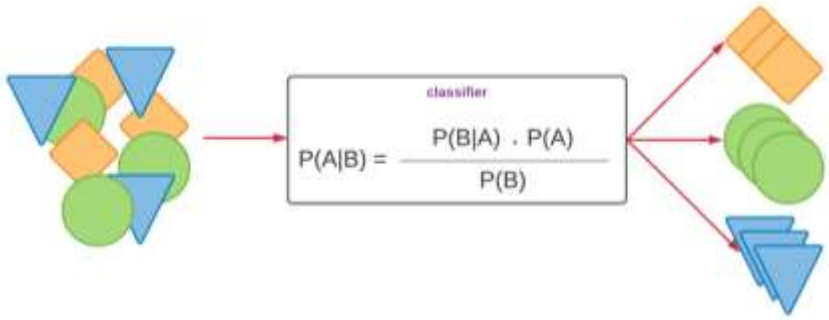
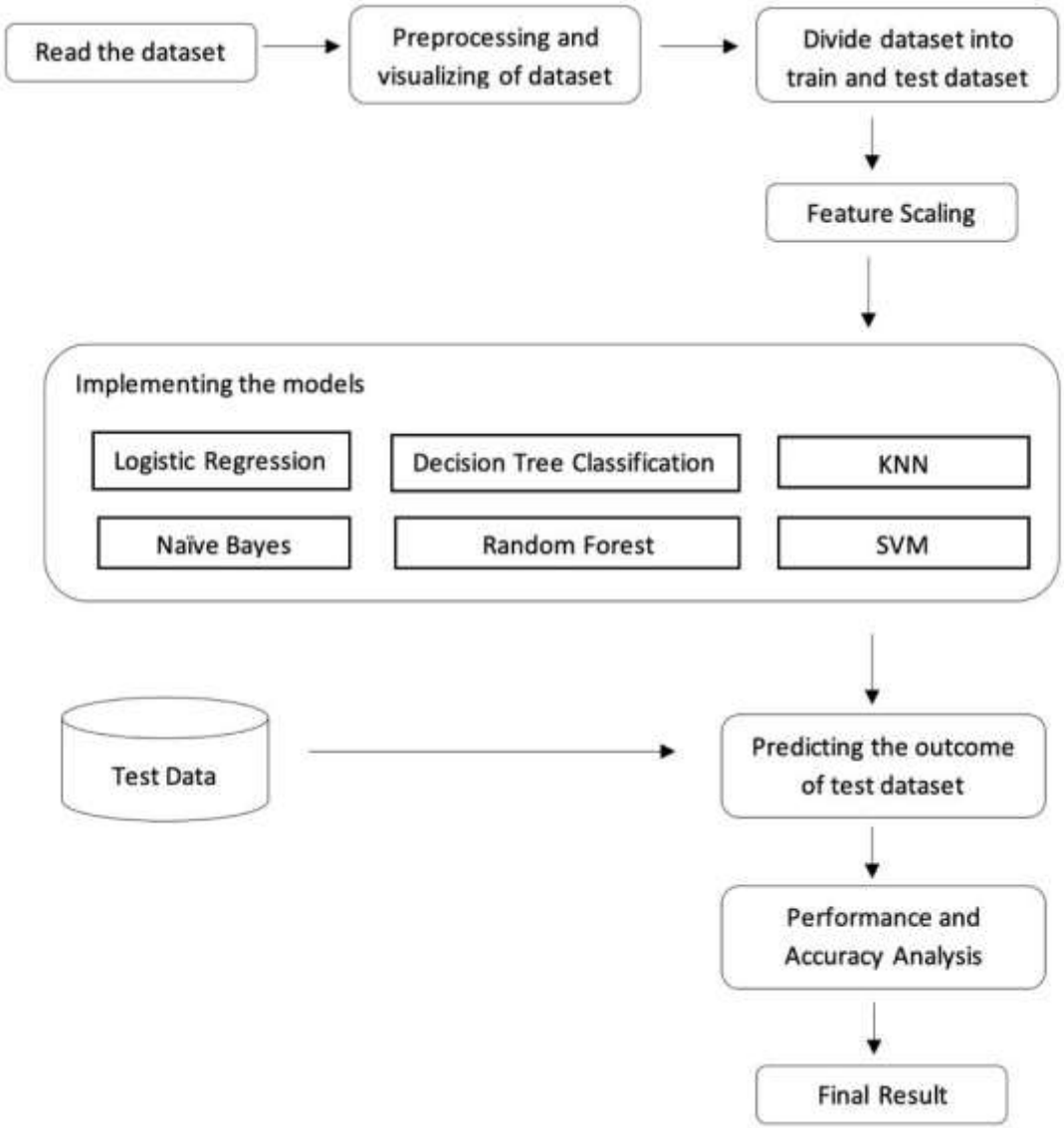


Figure 10: Naïve Bayes classifier

METHODOLOGY



SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

For the study, the first and foremost step was to acquire the required datasets and understand them. This involved data cleaning and pre-processing followed by data visualization for both the Iris and Wine Quality datasets. The datasets were visualized using various plots such as fractional percentage for each class, pair plots, correlation matrices and violin plots using python libraries of matplotlib and seaborn.

The visualization process provided some important observations about the datasets being used. Using this we collected the similarities, differences and some key points in the data.

The dataset was then split into train and test. Training set was 70% of the total dataset while Test data set was 30%. Training set was used as an input to prepare the model while Test set was used to test the accuracy.

This was followed by the training of all the six models that are the focus of this study, namely – K-Nearest Neighbors, Logistic Regression, Decision Tree classification, SVM, Random Forest & Naive Bayes. The models were trained on the training dataset. Then the models were fed the test data and the predicted values were analyzed for the results and calculation of the evaluation metrics.

A confusion matrix was plotted for each model and the evaluation metrics of accuracy, precision, recall and f1-score were calculated for each model on both the datasets.

RESULTS AND ANALYSIS

For the Iris dataset, we observed that the highest accuracy, 95.5%, was obtained for KNN and the least accuracy 88.88%, was obtained using Logistic Regression. The same has been tabulated and represented below for the models used.

Model	Accuracy
KNN	95.50%
Decision Tree	93.33%
Logistic Regression	88.88%
SVM	93.33%
Naïve Bayes	91.11%
Random Forest	91.11%

Table 3: Classification algorithms and their accuracies for iris dataset

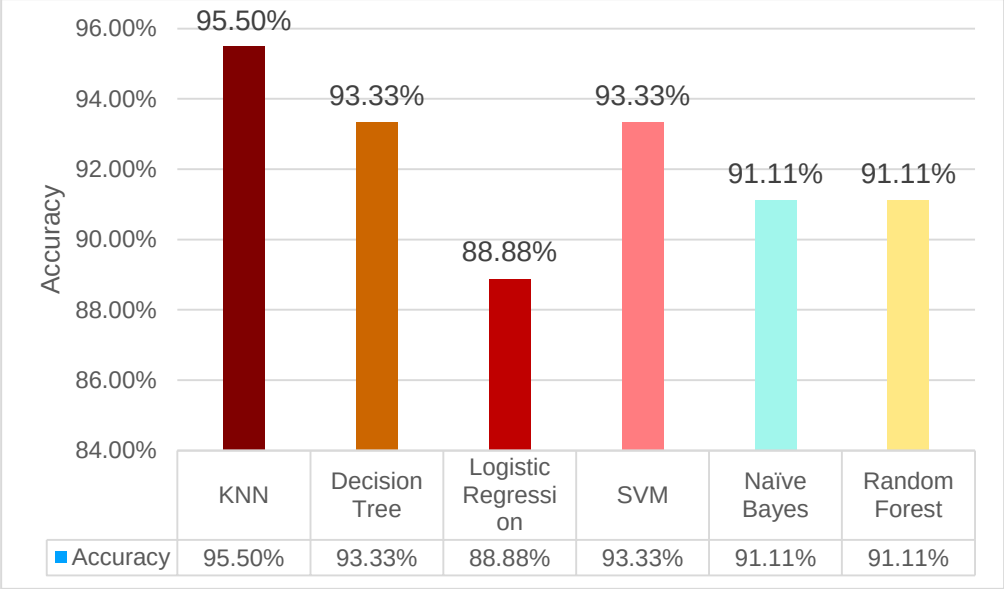


Figure 11: Plot of accuracies of different algorithms for iris dataset

For the Wine Quality dataset, we observed that the highest accuracy, 100%, was obtained for Decision tree and the least accuracy 82.29%, was obtained using KNN. The same has been tabulated and represented below for the models used.

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

Model	Accuracy
KNN	82.29%
Decision Tree	100.00%
Logistic Regression	98.33%
SVM	98.75%
Naïve Bayes	98.75%
Random Forest	94.79%

Table 4: Classification algorithms and their accuracies for Wine Quality dataset

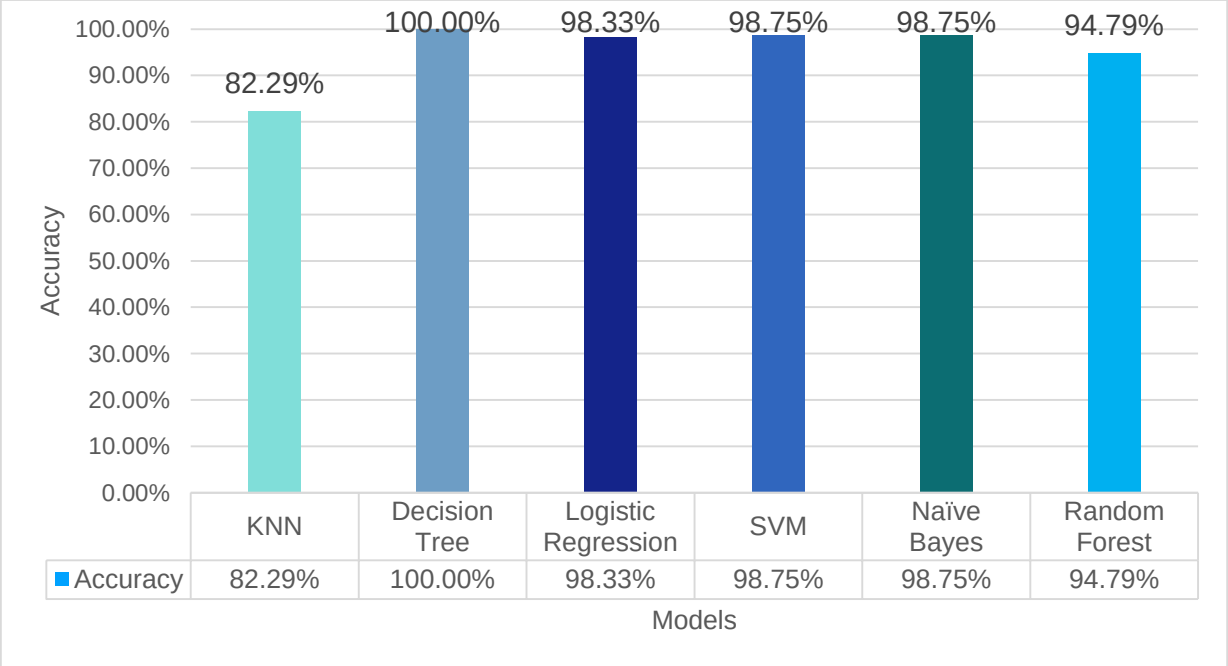


Figure 12: Plot of accuracies of different algorithms for Wine Quality dataset

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

Comparison of the accuracies of the models across the two datasets is depicted below:

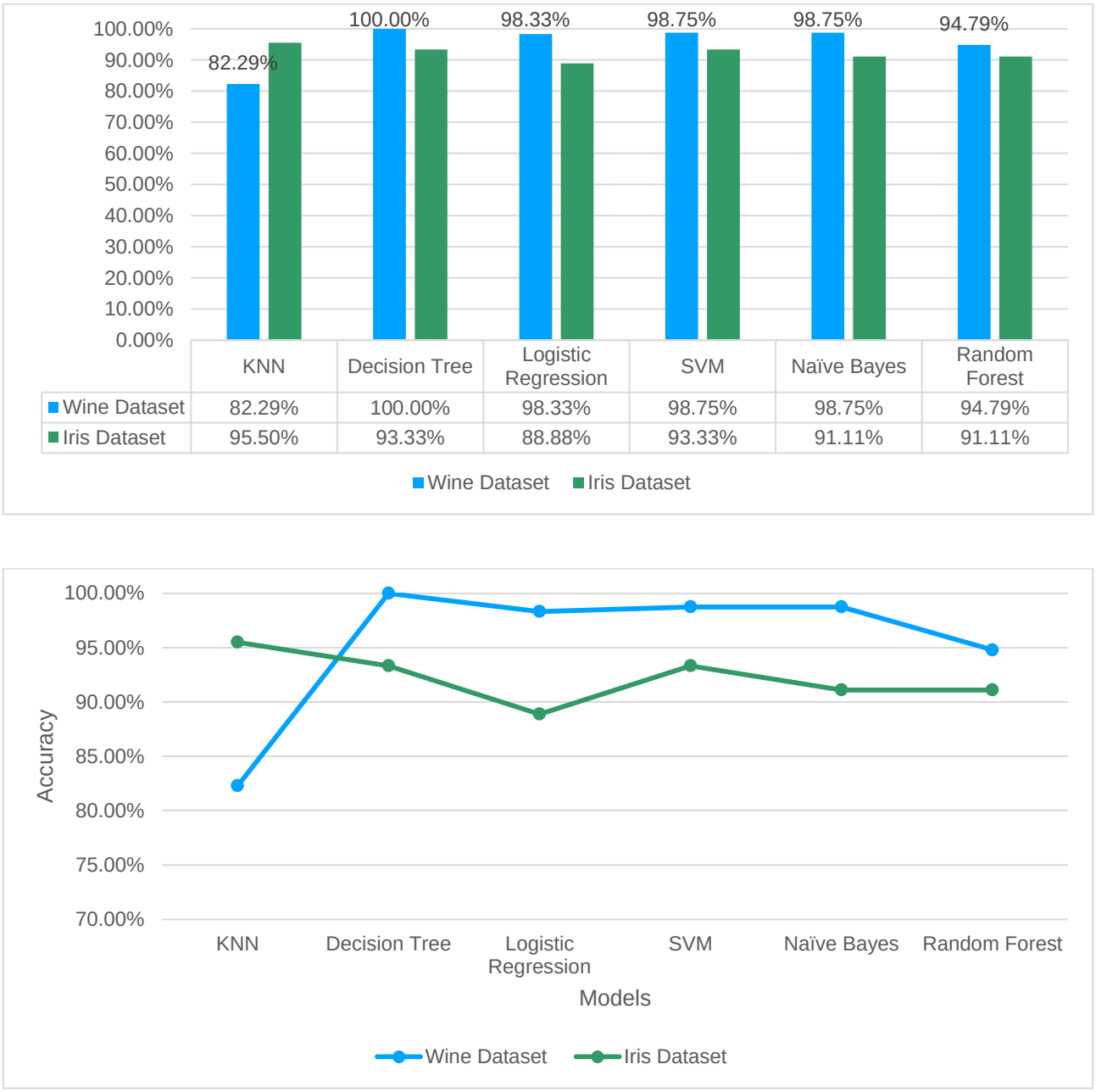


Figure 13: Plot of comparison of accuracies across different models between iris dataset and wine quality dataset

The evaluation metrics of precision, recall and f1 score, for both the models have been tabulated below:
For Iris dataset,

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

Model	Precision	Recall	F1-score
KNN	0.96	0.96	0.96
Decision Tree	0.93	0.91	0.91
Logistic Regression	0.92	0.89	0.89
SVM	0.95	0.93	0.94
Naïve Bayes	0.93	0.91	0.91
Random Forest	0.93	0.91	0.91

Table 5: Classification algorithm evaluation metrics for iris dataset

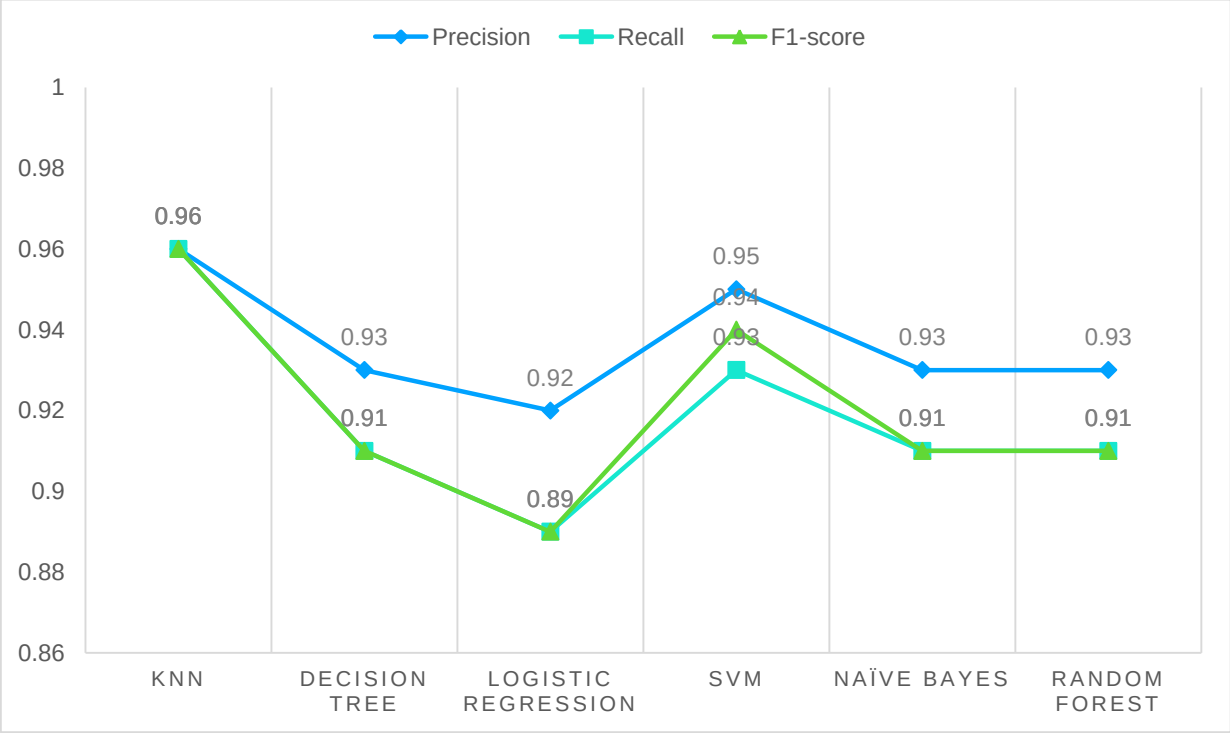


Figure 14: Plot of different evaluation metrics for classification algorithms on iris dataset

For Wine Quality dataset,

Model	Precision	Recall	F1-score
KNN	0.81	0.82	0.79
Decision Tree	1	1	1
Logistic Regression	0.98	0.98	0.98
SVM	0.99	0.99	0.99
Naïve Bayes	0.99	0.99	0.99
Random Forest	0.9	0.95	0.92

Table 6: Classification algorithm evaluation metrics for wine quality dataset

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

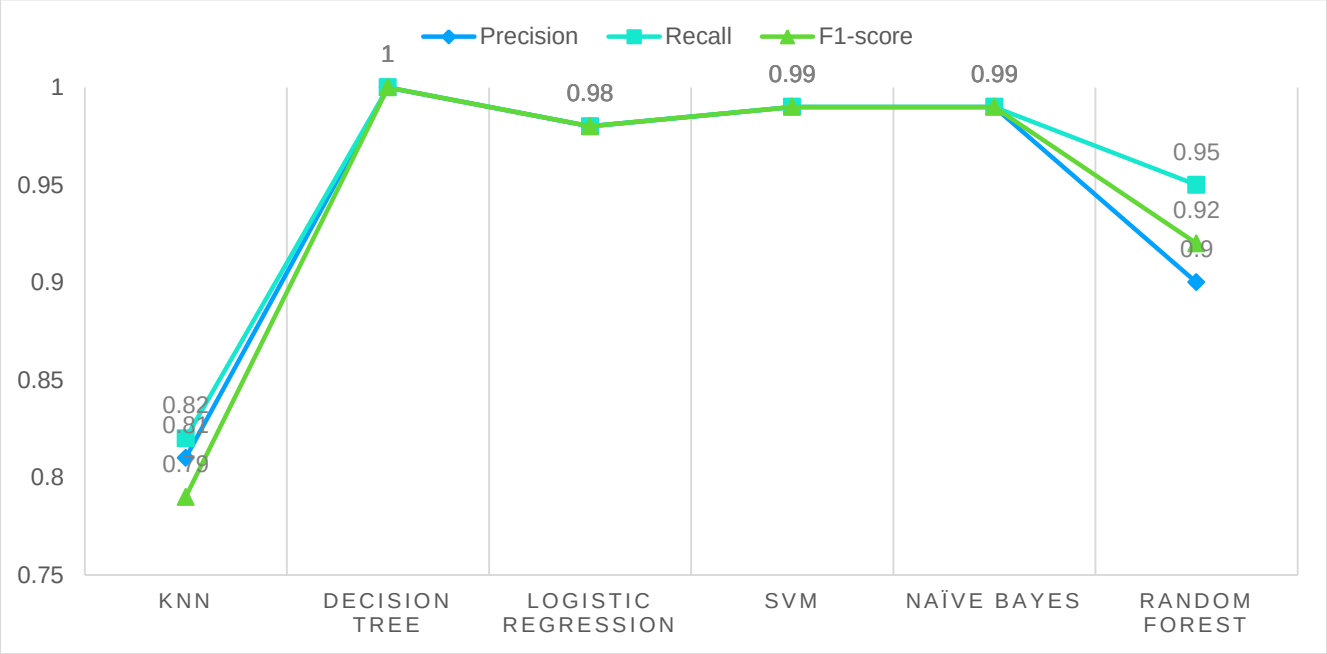


Figure 15: Plot of different evaluation metrics for classification algorithms on wine quality dataset

Conclusion and Future Scope

On further studying the results and comparing them with the database these inferences can be drawn

1. KNN is heavily influenced by the distribution of datapoints. If the datapoints are not well separated or overlapping it does not perform very well as seen in the case of Wine Quality Dataset whereas it performs very well if the datapoints are well separated even if only along a few parameters.
2. Decision Tree classifier works in a step-by-step approach when evaluating different parameters, which helps in the case of overlapping datapoints as observed in wine Quality Dataset.
3. One peculiar observation in our results was that decision Tree classifier performed better than Random Forest which can be potentially explained by the fact that wine quality dataset was not properly balanced and the records of Grade B far outnumbered the other grades. As Random Forest works on the principle of bootstrapped databases this could be a reason that the results were better in the case of decision tree classifier.

One of the most important decisions when trying to draw insights from any sort of data is selecting which algorithm to use. The results and efficiency highly depend on this choice. This document helps in classifying which algorithm works best for a simple classification-based datasets (iris dataset & wine quality in this case) and further more models and datasets can be included in this study to improve the classification and give insights to more people.

REFERENCES

1. *Kannapiran, T & Singh, Ajay & Rai, Prakhar & Gupta, Sachin. (2018). Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning. 1-4. 10.1109/CCAA.2018.8777643.*
2. *Prathima, P & Kumar, R. (2021) Comparison on Iris Dataset Using Classification Techniques", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.8, Issue 8, page no. ppc315-c319, August-2021,*
3. *Yuanyuan Wu , Jing He , Yimu Ji , Guangli Huang , Haichang Yao , Peng Zhang , 2019. Enhanced classification models for iris dataset. 7th International Conference on Information Technology and Quantitative Management*
4. *Alghobiri M. 2018. A Comparative Analysis of Classification Algorithms on Diverse Datasets: Engineering, Technology & Applied Science Research Vol. 8, No. 2, 2018, 2790-2795*
5. *Muhamedyev, R. & Yakunin, K. & Iskakov, S. & Sainova, S. & Abdilmanova, Ainur & Kuchin, Yan. (2015). Comparative analysis of classification algorithms. 96-101. 10.1109/ICAICT.2015.7338525.*
6. *Gupta Y. 2017. Selection of important features and predicting wine quality using machine learning techniques, 6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India*
7. *Gupta, M. Vanmathi, C. (2021) A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-10 Issue-1, May 2021*

SJIF Impact Factor 2022: 8.197| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 7 | Issue: 6 | June 2022 - Peer Reviewed Journal

8. *Sharma, N. (2018) Quality Prediction of Red Wine based on Different Feature Sets Using Machine Learning Techniques, International Journal of Science and Research (IJSR) ISSN: 2319-7064*
9. *Dahal, K. , Dahal, J. , Banjade, H. and Gaire, S. (2021) Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11, 278-289. doi: 10.4236/ojs.2021.112015.*
10. *Javidi, B., 2002. Image recognition and classification: algorithms, systems, and applications. CRC Press.*
11. *Mitchell, T.M., 2006. The discipline of machine learning (Vol. 3). Carnegie Mellon University, School of Computer Science, Machine Learning Department.*
12. *Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.*
13. *Domingos, P., 2012. A few useful things to know about machine learning. Communications of the ACM, 55(10), pp.78-87.*
14. *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.*