A Comparative Study of Classification Algorithms

Tanmay Kalani - S20160010096
B.Tech, CSE Dept.
IIITS
SriCity, India
tanmay.k16@iiits.in

G. Mary Ankita - S20160010029
B.Tech, CSE Dept.
IIITS
SriCity, India
maryankitha.g16@iiits.in

***Abstract*—Data Mining is the way toward recovering and distinguishing helpful data with astute algorithms from an informational collection and change data into fathomable shape. Classification is one of the data mining methods. It is a procedure of relegating substances to an officially characterized class by looking at features. Classifying data is the most well-known algorithm utilized for finding a mine standard from a substantial database. It is utilized to discover in which class every datum occurrence is connected inside data set. A near investigation of classification algorithms, for example, Decision Trees, Neural Networks and Naive Bayes Classifier has been finished. Objective of this investigation is to give an audit these algorithms. A general thought of Data Mining is classification talked about pursued by examination of algorithms. While considering these methodologies this paper gives a comprehensive review of various classification calculations and their features and impediments.**

***Index Terms*—Data Mining, Classification, Neural Network, Decision Tree, Naive Bayes Classification**

I. Introduction

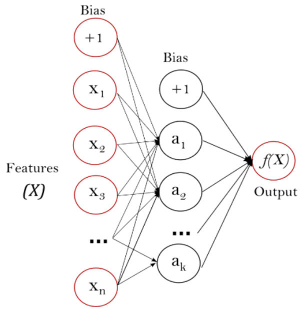
Data Mining is a procedure of extraction of helpful data from extensive measure of raw data. It is utilized to find important information and principles from data. Development of the classification criteria is finished utilizing a set of data instances for which related classes are known ahead of time. This sort of learning is named as **supervised learning**. Such algorithms are prepared for making choices in new circumstances.

Data sets utilized for training and comparing various algorithms are standard classification informational data sets by **scikit-learn general data set API**, namely **Iris**, **Breast Cancer** and **Wine** data sets and involve 150, 569 and 178 classified instance respectively. These are the data sets regularly utilized by the machine learning networks to benchmark algorithms which are utilized to take care of real-life problems.

A similar investigation of algorithms, for example, **Neural Networks**, **Bayes Classifier** and **Decision Trees** has been finished. **Scikit-learn** models are utilized to train, classify and measure precision and accuracy of data sets. Bayesian Classifier is **Gaussian Naive Bayes Classifier**. The Neural Network implemented is a **MLP(Multi-Layer Perceptron)** classifier. Decision trees chip away at the optimized adaptation of **CART** algorithm[1].

II. Information Theory

A. Neural Network



The neural network utilized here is a **Multi-Layer perceptron** which is a non-linear function approximator which can be utilized for classification . MLPClassifier learns using function

$$f(\cdot) : R^m \rightarrow R^o$$

where m and o are the dimensions of input and output respectively.

The neural node in the hidden layer transforms the values from the previous layer by a weighted sum $z = w_1x_1 + w_2x_2 + ... + w_nxn$ which is followed by a non-linear activation function - *hyperbolic tanh*

$$z = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

MLP training uses *Stochastic Gradient Descent*. It updates parameters using gradient of the loss function respect to parameter that requires adaptaion

$$w \leftarrow w - \eta(\alpha\frac{\partial R(w)}{\partial w} + \frac{\partial Loss}{\partial w})$$

where η is the learning rate, α is a on-negative hyper-parameter that controls the magnitude of penalty and $Loss$ is loss function.

MLP uses Square Loss Error function

$$Loss = \frac{1}{2}\sum_{r=1}^{c}(t_r - z_r)^2$$

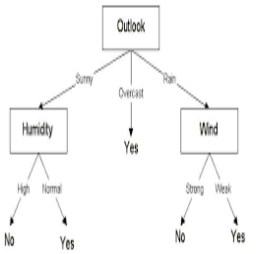
Time complexity of a network with h hidden layers and k nodes, i input nodes and o output nodes is

$$O(n \cdot m \cdot h^k \cdot o \cdot i)$$

where *n* is the number of training samples and *m* is the number of features.

Each of the data sets were trained with a neural network of one hidden layer with 100 nodes, $\alpha = 0.0001$, $\eta = 0.001$ over 200 epochs or more and stopping criteria $\epsilon = 0.01$.

B. Decision Trees



Decision Trees are a non-parametric supervised learning method used for classification. It makes twofold choices at any node i.e. it contructs a binary decision tree.

If a target is classification outcome taking on values 0,1,,K-1, for node *m*, representing a region R_m with N_m observations, let

$$p_{m,k} = 1/N_m \sum_{x_i \in R_m} I(y_i = k)$$

let be the proportion of class k observations in node m. Common Measures of impurity in **GINI** (variance impurity)

$$H(X_(m)) = \sum_k p_{mk}(1 - p_{mk})$$

and **Cross-Entropy** (information impurity)

$$H(X_m) = -\sum_k p_{mk} log(p_{mk})$$

and **Missclassification**

$$H(X_m) = 1 - max(p_{mk})$$

where X_m is the training data in node m.

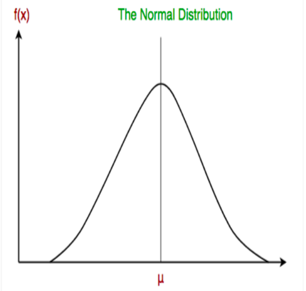
CART(Classification and Regression Trees) is extremely similar to **C4.5**[2][3], yet it varies in that it underpins numerical target variables (regression) and does not compute rule sets. CART constructs binary trees utilizing the feature and threshold that yield the largest information gain at every node.

The run time to construct a binary tree is $O(n_{samples} n_{features} log(n_{samples}))$ and query time is $O(log(n_{samples}))$. In order to find the features that offer the largest reduction in entropy there is a penalty of $O(n_{features})$ at each node. So the final cost adds up to $O(n_{features} n_{samples}^2 log(n_{samples}))$.

Every data set has been trained and tested on **GINI** and **Cross-entropy** impurity functions with different tree depths.

C. Gaussian Naive Bayes Classifier

Naive Bayes method are an arrangement of supervised learning algorithms dependent on applying Bayes' theorem with the "naive" suspicion of conditional independence between every pair of features given the estimation of the class variable. Whenever plotted, it gives a bell formed bend which is symmetric about the mean of the feature values as demonstrated as follows:



A Gaussian distribution is also called Normal distribution.

The likelihood of the features is assumed to be Gaussian:

$$P(X_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)}{2\sigma_y^2} \right)$$

Time complexity of Gaussian Naive Bayes is $O(nm)$ where n is the number of training data instances and m is the dimensionality of the features.

III. Training and Testing

Data sets utilized for training and comparing at the different algorithms are the standard CLASSIFICATION data sets by **scikit-learn** general informational index API, namely **Iris**, **Breast Cancer** and **Wine** data set and contain 150, 569 and 178 already classified instances individually.

A. Iris Data set

The Iris flower data set or Fisher's Iris data set is a multivariate data set presented by the British analyst and scholar Ronald Fisher[4]. This data includes 150 effectively classified instances with a class dispersion of 33.33%. Every datum instance comprises of 4 features (properties) in particular, sepal length, sepal width, petal length, petal width. Every one of these lengths are estimated in cms. Each set of these features is classified into either of the 3 classes to be specific $Setose$, $Versicolor$ and $Virginica$.

1) Neural Network:

- Ratio of test set to data set was set at 0.25
- Iris data set did not converge for 200 epochs, so maximum iterations was set at 1000.
- Correctnesses on training and testing data sets were 97% and 92% individually, which was bad enough. This happened on the grounds that information was not scaled. In this way, **data normalization** was executed as a **preprocessing** step with *StandardScaler* by *sklearn*.
- After standardization of data, accuracy on test data boosted to 97% as show in "Fig. 1".

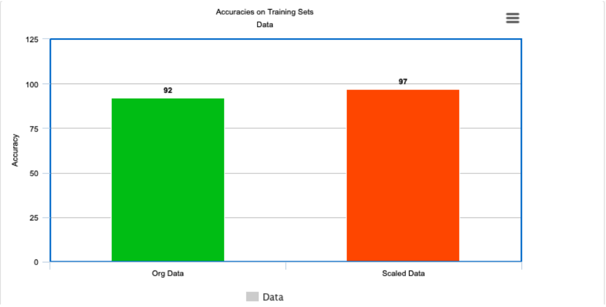


Fig. 1. Accuracy on training data sets

- Colorbar in ”Fig. 2” depicts weights of every one of the features .Rows in the plot are 4 feature names and the columns are 100 nodes of hidden layer. Blue is related with positive values and green is related with negative values. The more the blue zone in the strip of an element, the more is its significance. Notice the strip of $petal\ width$,which has the maximum number of blue focuses, hence is the central factor.

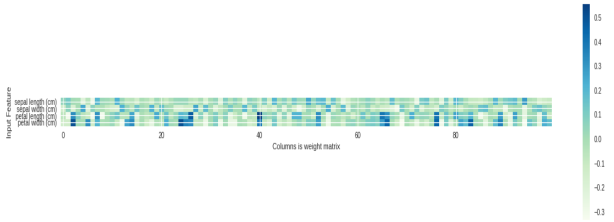


Fig. 2. Accuracy on training data sets

2) *Decision Trees:*

- Test Set was one-fourth the size of data set.
- At the point when the tree was permitted to develop completely with **Cross-Entropy** impurity, accuracy on the training set was 100%, which is an example of over-fitting. So **pre-pruning** techniques, for example, limiting max depth was applied. Subsequently, accuracy on the training set decreased from 100% to 99.107%, yet test set accuracy expanded from 92.105% to 94.30%.

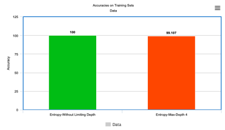


Fig. 3. Training Data set Accuracy with Entropy impurity

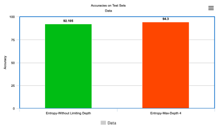


Fig. 4. Test Data set Accuracy with Entropy impurity

- Exactly when the tree was allowed to grow totally with **GINI** impurity, accuracy on the training set was 100%, which is an occurrence of over-fitting. So **pre-pruning** methods, for instance, limiting max depth was associated. Thus, accuracy on the training set diminished from 100% to 98.214%, yet test accuracy went down from 92.105% to 89.47%.

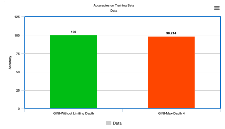


Fig. 5. Training Data set Accuracy with GINI impurity

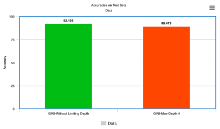


Fig. 6. Test Data set Accuracy with GINI impurity

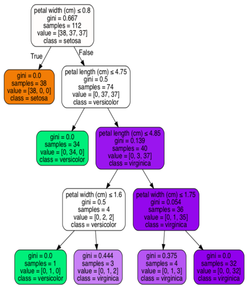


Fig. 7. Decision Tree with GINI impurity with limited max-depth

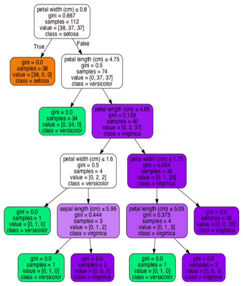


Fig. 8. Decision Tree with GINI impurity without limited max-depth

- ”Fig. 7” is a decision tree with a depth-limit of 4 and ”Fig. 8” is a decision tree with no depth restrain. Constraining a decision tree is a piece of $pre-pruning$ so model does not overfit and can classify new feature sets with a considerably higher accuracy.
- ”Fig. 9” demonstrates the feature significance. It tends to be inferred $petal\ width$ has the most elevated feature importance and is also at the $root$ of the Decision Tree. In this way, it assumes a major job in deciding class of the data instance.

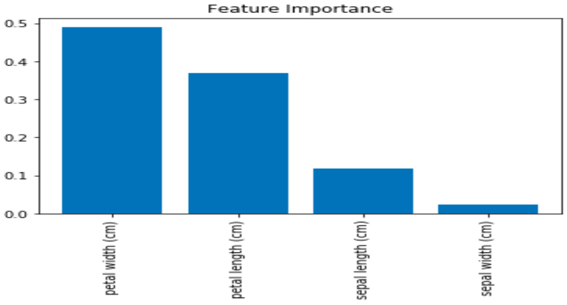


Fig. 9. Feature Importance of Iris Data Set

3) *Gaussian Naive Bayes:*

- Proportion of test set to data set was set at 0.25
- Accuracy of Gaussian Naive Bayes was 97.368%.

B. *Breast Cancer Data Set*

Features in the data set are computed from a digitized image of a **fine needle aspirate** of a breast mass. They describe characteristics of the cell nuclei present in the image and was first used in 1993 by W.N. Street[5]. The data set comprises of 569 already classified instances. Each data instance consisted

TABLE I
IRIS DATA SET

Algorithm	Accuracy
Multi-Layer Perceptron	97%
Decision Tree - Cross Entropy	94.30%
Decision Tree - GINI Impurity	89.47%
Gaussian Naive Bayes	97.368%

of 30 features (attributes). Data instances are classified into two classes $Malignant$ and $Benign$ with 212 and 357 data instances respectively.

1) Neural Network:

- Ratio of test set to data set was set at 0.25
- Cancer data set did not converge for 200 epochs, so maximum iterations was set at 1000.
- Accuracies on training and testing data set were 93.89% and 92.30% respectively, which was not good enough. This happened because data was not scaled. So, **text normalization** was performed as a preprocessing step using $StandardScaler$ by $sklearn$.
- After standardization of data, accuracy on test data boosted to 95.804% as show in ”Fig. 10”.

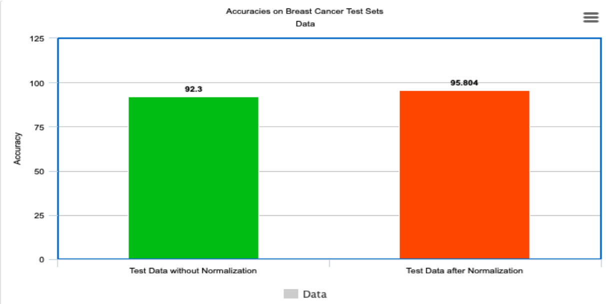


Fig. 10. Accuracy on test data sets

- Colorbar in ”Fig. 11” weights of each of the features. Rows in the plot are 30 feature names and the columns are 100 nodes of hidden layer. Blue is associated with a more positive value and green is associated with negative values. The more the blue area in the row of a feature, the more is its importance. Notice the strips of $Smoothness$ $error$ and $Fractal$ $Dimension$ $error$ which have all green pixels, thus, they do not play a significant role in classification whereas $worst$ $radius$ has the maximum number of blue pixels which increases its importance.

2) Decision Trees:

- Test Set was one-fourth the size of data set.
- When the tree was allowed to grow fully with **Cross-Entropy** impurity, accuracy on the training set was 100%, which is a case of over-fitting. So **pre-pruning** methods such as $limiting$ max $depth$ was applied. As a result, accuracy on the training set reduced from 100% to 98.591%, but test accuracy increased from 94.405% to 95.804%.

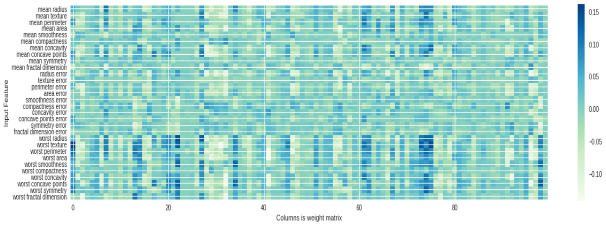


Fig. 11. Color Bar of features

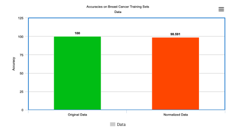


Fig. 12. Training Data set Accuracy with Entropy impurity

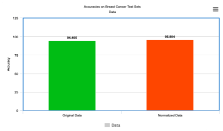


Fig. 13. Test Data set Accuracy with Entropy impurity

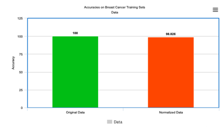


Fig. 14. Training Data set Accuracy with GINI impurity

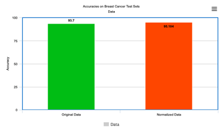


Fig. 15. Test Data set Accuracy with GINI impurity

- When the tree was allowed to grow fully with **GINI** impurity, accuracy on the training set was 100%, which is a case of over-fitting. So **pre-pruning** methods such as $limiting$ max $depth$ was applied. As a result, accuracy on the training set reduced from 100% to 98.826%, but test accuracy increased from 93.706% to 95.104%.

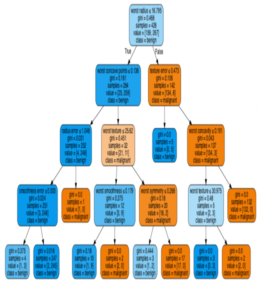


Fig. 16. Decision Tree with GINI impurity with limited max-depth

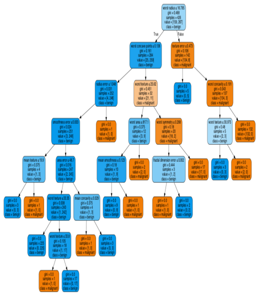


Fig. 17. Decision Tree without GINI impurity without limited max-depth

- ”Fig. 16” is a decision tree with a depth-limit of 4 and ”Fig. 17” is a decision tree without any depth-limit. Limiting a decision tree is a part of pre-pruning so that model does not overfit and can classify new data with a much higher accuracy.
- ”Fig. 18” shows the feature importance. It can be inferred $worst$ $radius$ has the highest feature importance and is also at the $root$ of the Decision Tree. Thus, it plays a major role in deciding class of the data instance. Worst Radius may be an important feature, but it might not tell

TABLE II
BREAST CANCER DATA SET

Algorithm	Accuracy
Multi-Layer Perceptron	95.804%
Decision Tree - Cross Entropy	95.804%
Decision Tree - GINI Impurity	95.104%
Gaussian Naive Bayes	92.30%

us that a higher radius does not indicate sample being $Malignant$ or $benign$.

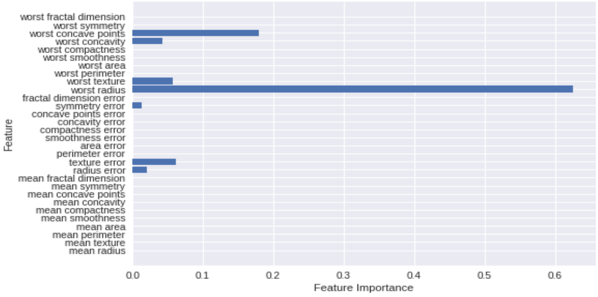


Fig. 18. Feature Importance of Breast Cancer Data Set

3) Gaussian Naive Bayes:

- Proportion of test set to data set was set at 0.25
- Accuracy of Gaussian Naive Bayes was 92.30%.

C. Wine Data Set

The data is eventual outcome of manufactured examination of wines grown in same district of Italy by 3 particular cultivators. The data set contains 178 viably classified data events. Each of these instanced contain 13 features(attributes). Each of these feature sets are classified into three classes namely $class_0$, $class_1$,$class_2$ with 59, 71, 48 independently.

1) Neural Network:

- Ratio of test set to data set was set at 0.25
- Wine data set did not converge for 200 epochs, so maximum iterations was set at 1000.
- Correctnesses on training and testing data set were 60.90% and 68.88% independently, which was terrible enough. This occurred in light of the fact that data was not scaled. Along these lines, **data normalization** was executed as a *pre-processing* step using *Standard Scaler* by *sklearn*.
- After standardization of data, accuracy on test data boosted to 100% as show in "Fig. 19".
- Colorbar in "Fig. 20" weights of all of the features. Rows in the plot are 13 feature names and the columns are 100 nodes of hidden layer. Blue is connected with a more positive regard and green is connected with negative regard. The more the blue zone in the strip of a feature, the more is its importance. Notice the strips of $OD280/OD315$ has the most outrageous number of blue pixels which constructs its criticalness.

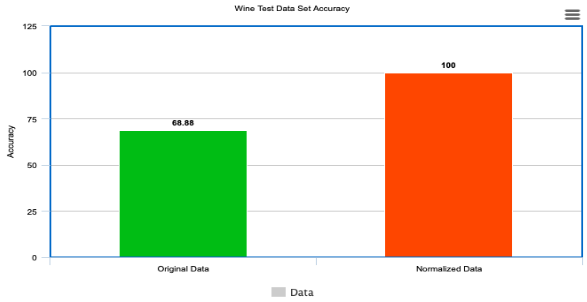


Fig. 19. Accuracy on test data sets

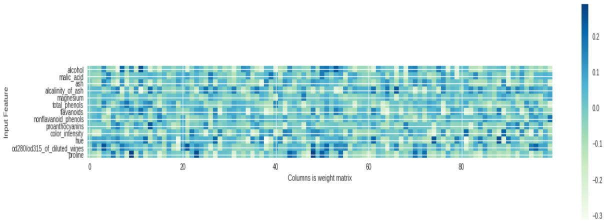


Fig. 20. Color Bar of features

2) Decision Trees:

- Test Set was one-fourth the size of data set.
- Right when the tree was allowed to grow totally with Cross-Entropy impurity impact, precision on the training set was 100%,which is a case of over-fitting. So pre-pruning methods such as limiting max depth was associated. Hence ,accuracy on the arrangement set lessened from 100% to 98.496%, yet test precision extended from 93.33% to 94.33%.

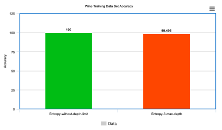


Fig. 21. Training Data set Accuracy with Entropy impurity

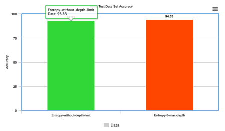


Fig. 22. Test Data set Accuracy with Entropy impurity

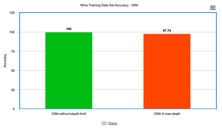


Fig. 23. Training Data set Accuracy with GINI impurity

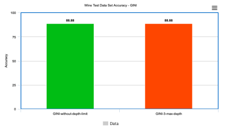


Fig. 24. Test Data set Accuracy with GINI impurity

- When the tree was allowed to grow fully with **GINI** impurity, accuracy on the training set was 100%, which is a case of over-fitting. So **pre-pruning** methods such as *limiting max depth* was applied. As a result, accuracy on the training set reduced from 100% to 97.74%, but test accuracy remained constant at 88.88%.
- "Fig. 25" is a decision tree with a depth-limit of 3 and "Fig. 26" is a decision tree without any depth-limit.

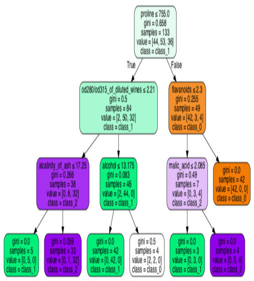


Fig. 25. Decision Tree with GINI impurity with limited max-depth

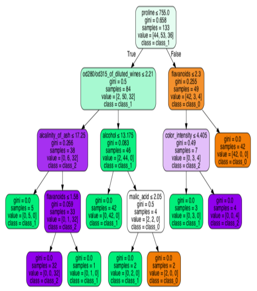


Fig. 26. Decision Tree without GINI impurity without limited max-depth

TABLE III
WINE DATA SET

Algorithm	Accuracy
Multi-Layer Perceptron	100%
Decision Tree - Cross Entropy	94.33%
Decision Tree - GINI Impurity	97.74%
Gaussian Naive Bayes	95.55%

Limiting a decision tree is a part of pre-pruning so that model does not overfit and can classify new data with a much higher accuracy.

- ”Fig. 27” exhibits the segment note worthiness. It might be inferred *proline* has the most raised component importance and is in like manner at the root of the Decision Tree. As such, it plays a critical role in picking class of the data instance.

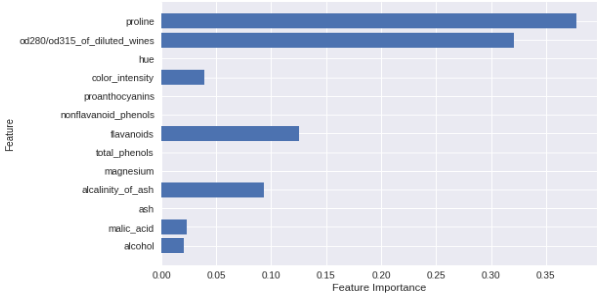


Fig. 27. Feature Importance of Wine Data Set

3) Gaussian Naive Bayes:

- Proportion of test set to data set was set at 0.25
- Accuracy of Gaussian Naive Bayes was 95.55%.

IV. Conclusion

All classifiers were trained and testes on same dissemination of training and testing data sets. Every one of them gave palatable outcomes considering they were prepared with not very many data instances.

Gaussian Naive Bayes scored the highest accuracy (97.368%) for IRIS data set. Decision Tree with Cross-Entropy impurity scored highest accuracy (95.804%) for Brest Cancer Data set. Decision Tree with GINI impurity scored the highest accuracy (95.55%) for Wine Data set.

Gaussian Naive Bayes achieved the best appealing results for all of these data sets. Decision Tree Classifier had the humblest time complexity. Neural Networks can have high accuracy for yet they take the longest to execute and have extensibility issues due to their to an large and complex nature.

Bayesian Classifier is a probabilistic model and can have certain stable states. To work with continous data, a binning algorithm is utilized however in the event that not utilized appropriately there can happen loss of important data.

On the off chance that the Decision Trees are permitted to grow totally they may frame an over-fitted model, which is useful for training instances however probably won’t perform useful for new test data. Besides, these trees are base on heuristic calculations, for example, greedy algorithms where locally ideal choices are made at every node which does not ensure all around optimal decision tree.

Neural Networks utilizes a non-linear activation function. It has a capacity to learn non-linear models at the same time, if the beginning arbitrary weights are not choosen appropriately it can wind up in a local minimum least with different validation accuracy. Likewise, it is imperative to include scaling.

References

[1] https://scikit-learn.org/stable/modules/tree.html

[2] Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI, Mohammed ERRITALI, A comparative study of decision tree ID3 and C4.5, 2014,

[3] I. S. Jacobs and C. P. Bean, “ REVIEW OF DECISION TREE DATA MINING ALGORITHMS: ID3 AND C4.5”, 2015.

[4] Fisher, R.A. The use of multiple measurements in taxonomic problems Annual Eugenics, 7, Part II, 179-188 (1936); also in Contributions to Mathematical Statistics (John Wiley, NY, 1950).

[5] W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IST/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

[6] M. L. Zhang and Z. H. Zhou. A Review on Multi-Label Learning Algorithms. In: IEEE Transactions on Knowledge and Data Engineering 26.8 (Aug. 2014), pp. 1819 1837. issn: 1041-4347. doi: 10.1109/TKDE.2013.39.