DATA MINING HOMEWORK- 1

DhayaliniNagaraj

A20359686

The computational process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data is known as Data Mining.

There are two forms of data analysis that is used for extracting models describing important classes or to predict future data trends. These two forms are as follows

- Classification
- Prediction

In Data Mining, classification is the process of assigning predetermined categories to an item or an instance that has not been encountered. The categories are assigned based on the knowledge of categories of the similar items that we have dealt with.

In this assignment, the datasets Iris, Vote and Diabetes datasets are used to do data analysis and the classifications using different decision tree algorithms.

Data Analysis

Iris dataset

Number of instance: 150
Number of attributes: 5 (4 Numerical + 1 Nominal)
*Types of attributes:*Numerical and Nominal.

The Iris dataset has data about 150 instances.It is gathered across 4 numerical attributes or traits, with the class attribute of type nominal.

Numeric Attributes:

S. No	Attributes	Minimum Value	Maximum Value	Mean	Standard Deviation
1	Sepallength	4.3	7.9	5.843	0.828
2	Sepalwidth	2	4.4	3.054	0.434
3	Petallength	1	6.9	3.759	1.764
4	Petalwidth	0.1	2.5	1.199	0.763

Nominal Attribute:

S. No	Attribute	Values	Count
5	Class	Iris-setosa	50
		Iris-versicolor	50
		Iris-virginica	50

Vote Dataset

Number of instances: 435
Number of attributes: 17 (17 Nominal)
Types of attributes: Nominal
The Vote dataset has vote details of the Congressmen in the U.S House of representatives. It has 435 instances. There are 16 attributes and a class attribute which is 17 in total. All the attributes are of type Nominal.

Diabetes Dataset

Number of instances: 768
Number of Attributes: 9 (8 Numerical + 1 Nominal)
*Types of attributes:*Numerical and Nominal.

The diabetes dataset is a collection of data across 8 attributes of numerical data type and a class attribute of nominal type. The diabetes risk factors represent 768 feminine instances of Pima Indian origin.

Iris dataset

Number of instance: 150
Number of attributes: 5 (4 Numerical + 1 Nominal)
*Types of attributes:*Numerical and Nominal.

Numeric Attributes:

S. No	Attributes	Minimum Value	Maximum Value	Mean	Standard Deviation
1	Sepallength	4.3	7.9	5.843	0.828
2	Sepalwidth	2	4.4	3.054	0.434
3	Petallength	1	6.9	3.759	1.764
4	Petalwidth	0.1	2.5	1.199	0.763

Nominal Attribute:

S. No	Attribute	Values	Count
5	Class	Iris-setosa	50
		Iris-versicolor	50
		Iris-virginica	50

For a given data, we can find the range of each attribute which means the dispersion or spread of a set of values. Such measure indicates if the attribute values are widely spread out or if they are relatively concentrated around a single point such as the mean.

The measure of spread is the range and it can be defined as given an attribute x with a set of m values then,

Range(x)= max(x)- min(x)

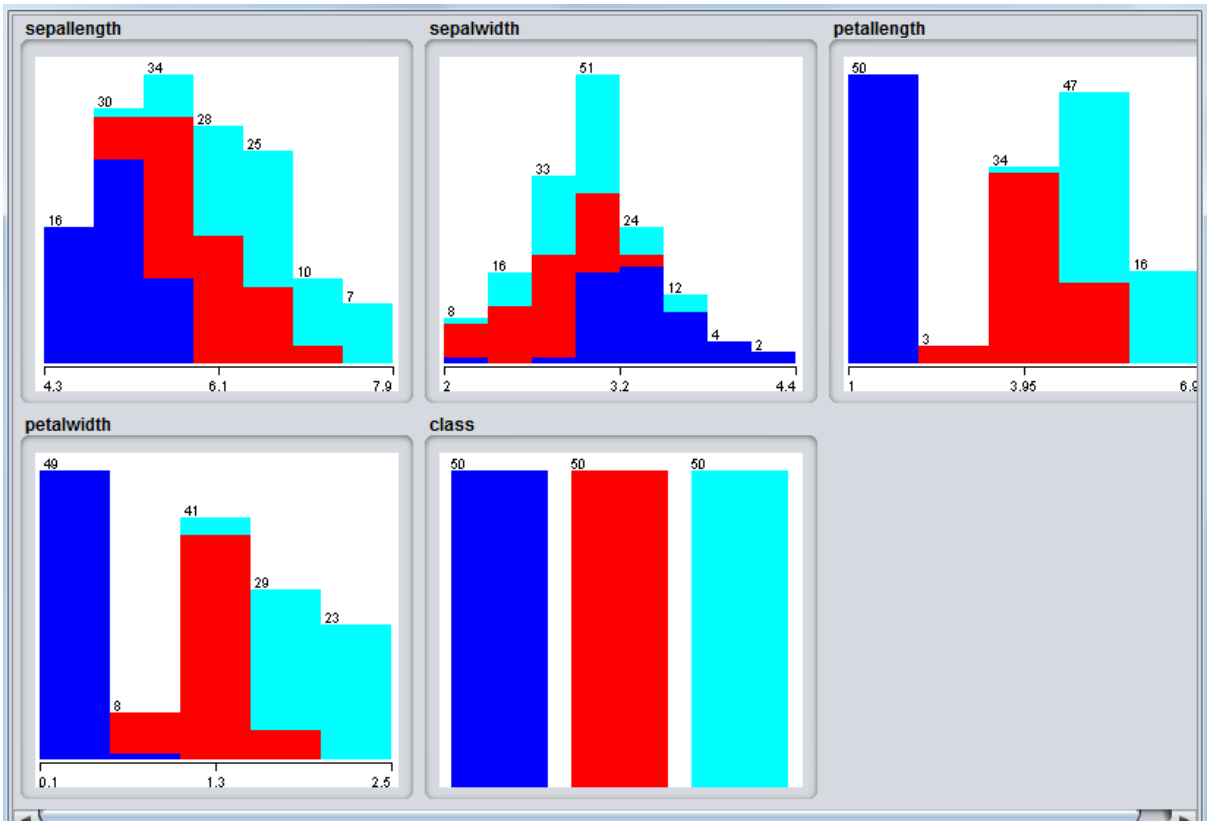
So the range of attributes of Iris dataset is

Attribute	Range
Sepallength	3.6
Sepalwidth	2.4
Petallenght	5.9
Petalwidth	2.4

Although the range identifies the maximum spread, it can be misleading if most of the values are concentrated in a narrow band of values, but there are also a relatively small number of more extreme values.

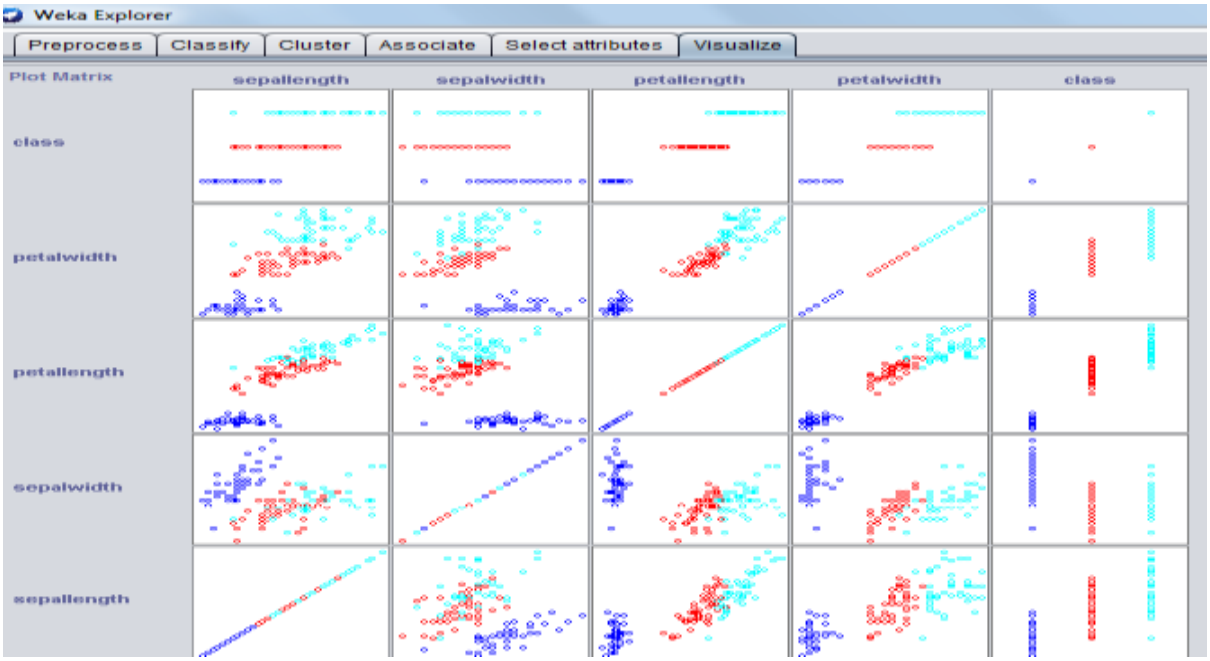
For example:
The range of each attribute may vary with the scale.Some attributes scale range may be from 0 to 10 whereas some may vary from 1000 to 2000. In that case the classification made from the attributes range may mislead us.

In order to avoid this, we can use normalization.



Using Weka tool, we can Visualise the Iris data. The image shows us a histogram for the attribute distribution for single selected attribute at a time, by default this is the class attribute.

The individual colours indicate the individual classes in the dataset.



From this scatterplots, number of things can be studied. Different colours represent different class. From this plots when we consider petalwidth or petal length attributes, it is easy for us to differentiate the class. But in Sepalwidth or sepallength, it is harder to separate the clusters by color. Based on the data analysis and the visualisation, it is difficult to classify the data with sepallength and sepalwidth because the overlapping is more and the classification will be impure.

Petallength attribute is the best attribute to classify the Iris data set because the overlapping is less and each class can be identified correctly without more confusion. The classification will not be impure when compared to classifications based on other attributes.

In weka tool, using attribute selector and different classifier, it was noticed that the classification is done correctly based on petal length attribute.

```
Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.GreedyStepwise -T -1.7976931348623157E308 -N -1 -num-slots 1
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Evaluation mode:    10-fold cross-validation


=== Attribute selection 10 fold cross-validation seed: 1 ===

number of folds (%)  attribute
          0(  0 %)    1 sepallength
         10(100 %)    3 petallength
          0(  0 %)    4 petalwidth
         10(100 %)    5 class
```

2. Decision Tree Algorithms

Decision tree, which is a classification technique, works by iteratively checking the attributes of the dataset for conditions at each level, based on which, the instances are grouped into several nodes, ideally upto a point at which each instance can be assigned a single definite label. There are several ways to do this, the mostly used method is built upon the degree of class impurity levels of nodes. The common class impurity measures are Entropy, Gini and Classification error.

There are several types of nodes in a decision tree, which we need to get accustomed to before delving deeper.

• Root node: This is a node which holds one or many outgoing edges with no incoming edge, for it is the origin of any decision tree.

• Internal node: This node possesses a single incoming edge and one or many outgoing edges. Internal node is used in propagating the tree from root node to the leaf nodes.

• Leaf node: These nodes are the ending points of a decision tree with one or many incoming edges and no outgoing edge.

Decision Stump Algorithm:

This algorithm produces a Decision Stump, which is a variant of decision tree where there is only one level. This implies that the size of a decision tree (the number of nodes in a tree) is always 3 and the internal nodes are absent. There is a single root node, as with all other decision trees, where the attribute with maximal information gain is checked for a condition based on the result of which instances are split into two groups and different class labels are assigned to them. In case of an instance missing the value of the selected attribute, it is tagged the class label that is assigned to the instances that satisfy the test condition.

Parameters:

There are no parameters affects this algorithm as the result is a one-level tree. The only parameter that Weka provides for Decision Stump is the 'Debug' parameter, which when set to True will output additional information to the console at some cases.

J48 Algorithm:

J48 decision tree algorithm is slightly modified form C4.5 in WEKA. For the given data-set it generates a classification-decision tree by recursive partitioning of data. Using Depth-first strategy the decision is grown. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests involving every distinct values of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy

gains of the binary cut based on each distinct values are calculated in one scan of 4 the sorted data. This process is repeated for each continuous attributes.

Parameters:

seed -- The seed used for randomizing the data when reduced-error pruning is used.

unpruned -- Whether pruning is performed.

confidenceFactor -- The confidence factor used for pruning (smaller values incur more pruning).

numFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.

numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model.

batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size.

reducedErrorPruning -- Whether reduced-error pruning is used instead of C.4.5 pruning.

useLaplace -- Whether counts at leaves are smoothed based on Laplace.

doNotMakeSplitPointActualValue -- If true, the split point is not relocated to an actual data value. This can yield substantial speed-ups for large datasets with numeric attributes.

debug -- If set to true, classifier may output additional info to the console.

subtreeRaising -- Whether to consider the subtree raising operation when pruning.

saveInstanceData -- Whether to save the training data for visualization.

binarySplits -- Whether to use binary splits on nominal attributes when building the trees.

doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime).

minNumObj -- The minimum number of instances per leaf.

useMDLcorrection -- Whether MDL correction is used when finding splits on numeric attributes.

collapseTree -- Whether parts are removed that do not reduce training error.

minNumObj, Unpruned, reduced error pruning, numfolds, confidence factor, seed are the important parameters. Because these are the parameters that affect the accuracy level and error level, time taken to classify the data are dependent on these parameters.

Random Forest Algorithm:

Random Forest is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random forest generally exhibits a substantial performance improvement over the single tree classifier such as CART and C4.5. It yields generalization error rate that compares favorably to Adaboost, yet is more robust to noise.

Parameters:

seed -- The random number seed to be used.

representCopiesUsingWeights -- Whether to represent copies of instances using weights rather than explicitly.

storeOutOfBagPredictions -- Whether to store the out-of-bag predictions.

numExecutionSlots -- The number of execution slots (threads) to use for constructing the ensemble.

bagSizePercent -- Size of each bag, as a percentage of the training set size.

numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model.

batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size.

printClassifiers -- Print the individual classifiers in the output

numIterations -- The number of iterations to be performed.

debug -- If set to true, classifier may output additional info to the console.

outputOutOfBagComplexityStatistics -- Whether to output complexity-based statistics when out-of-bag evaluation is performed.

classifier -- The base classifier to be used.

breakTiesRandomly -- Break ties randomly when several attributes look equally good.

doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime).

maxDepth -- The maximum depth of the tree, 0 for unlimited.

calcOutOfBag -- Whether the out-of-bag error is calculated.

numFeatures -- Sets the number of randomly chosen attributes. If 0, int(log_2(#predictors) + 1) is used.

maxDepth,numIterations, seed, breakTiesRandomly, numFeatures are the important attributes in this classifier. These parameters affect the time, accuracy, error level of the classifier.

Class Distribution:

Dataset	Class	Distribution
Iris Dataset	Iris-setosa	33.33%
	Iris-versicolor	33.33%
	Iris-virginica	33.33%
Vote Dataset	democrat	61.38%
	republican	38.62%
Diabetes Dataset	tested_negative	68.105%
	tested_positive	34.895%

Let us compare the distribution details with our results to arrive at a new set of inferences on the characteristics of the decision trees.

DECISION STUMP:

IRIS DATASET

Decision Stump	10 Cross Validation	Training Set
Correctly Classified Instance	100	100
Incorrectly classified Instance	50	50
Correctly classified Percentage	66.6667%	66.6667%
Incorrectly classified Percentage	33.3333%	33.3333%
Relative absolute error	50%	50%
Root relative squared error	70.7107%	70.7107%
Total Number of instances	150	150

Decision Stump	Percentage Split 20%	Percentage Split 50%	Percentage Split 66%	Percentage Split 90
Correctly Classified Instance	78	48	100	10
Incorrectly classified Instance	42	27	50	5
Correctly classified Percentage	65%	64%	66.6667%	66.6667%
Incorrectly classified Percentage	35%	36%	33.3333%	33.3333%
Relative absolute error	51.6379%	52.8316%	50%	50%
Root relative squared error	71.1443%	72.4626%	70.7107%	70.7107%
Total Number of instances	120	75	150	15

VOTE DATA SET:

Decision Stump	10 Cross Validation	Training Set
Correctly Classified Instance	416	416
Incorrectly classified Instance	19	19
Correctly classified Percentage	95.6322	95.6322
Incorrectly classified Percentage	4.3678	4.3678
Relative absolute error	16.693%	16.5385
Root relative squared error	41.2142%	40.6726
Total Number of instances	435	435

Decision Stump	Percentage Split 20%	Percentage Split 50%	Percentage Split 66%	Percentage Split 90
Correctly Classified Instance	334	210	143	39
Incorrectly classified Instance	14	7	5	4
Correctly classified Percentage	95.977%	96.7742%	96.6216%	90.6977%
Incorrectly classified Percentage	4.023%	3.2258%	3.3784%	9.3023%
Relative absolute error	17.8318%	17.2967%	16.4791%	27.1839
Root relative squared error	39.7676%	36.7968%	37.8685%	64.6536
Total Number of instances	348	217	148	43

DIABETES DATASET:

Decision Stump	10 Cross Validation	Training Set
Correctly Classified Instance	552	565
Incorrectly classified Instance	216	203
Correctly classified Percentage	71.875%	73.5677%
Incorrectly classified Percentage	28.125%	26.4323%
Relative absolute error	83.6467%	81.8217%
Root relative squared error	92.6909%	90.4671%
Total Number of instances	768	768

Decision Stump	Percentage Split 20%	Percentage Split 50%	Percentage Split 66%	Percentage Split 90
Correctly Classified Instance	403	282	202	59
Incorrectly classified Instance	211	102	59	18
Correctly classified Percentage	65.6352%	73.4375%	77.3946%	76.6234%
Incorrectly classified Percentage	34.3648%	26.5625%	22.6054%	23.3766%
Relative absolute error	88.9497%	82.27%	82.9438%	79.0404%
Root relative squared error	97.3455%	89.7266%	90.1142%	86.121%
Total Number of	614	384	261	77

instances				

J48 DECISION TREE

IRIS DATASET

J48	10 Cross Validation	Training Set	Percentage Split 20%	Percentage Split 66%	Percentage split 90%
Time Taken	0.03 sec	0.01	0.01	0.01	0.03
Number of Leaves	5	5	5	5	5
Size of Tree	9	9	9	9	9
Correctly Classified Instance	144	147	115	49	15
Incorrectly classified Instance	6	3	5	2	0
Correctly classified Percentage	96%	98%	95.8333%	96.0784%	100%
Incorrectly classified Percentage	4%	2%	4.1667%	3.9214%	0%
Relative absolute error	7.8705%	5.2482%	10.6933%	8.8979	2.1152%
Root relative squared error	33.6353%	22.9089%	36.0709%	33.4091	3.1533%
Total Number of instances	150	150	120	51	15

VOTE DATASET

J48	10 Cross Validation	Training Set	Percentage Split 20%	Percentage Split 66%	Percentage split 90%
Time Taken	0.01sec	0.	0.01	0.05	0.01
Number of Leaves	6	6	6	6	6
Size of Tree	11	11	11	11	11
Correctly Classified Instance	419	423	328	144	41
Incorrectly classified Instance	16	12	20	4	2
Correctly classified Percentage	96.3218	97.2414%	94.2529%	97.2973%	95.3488%
Incorrectly classified Percentage	3.6782%	2.7586%	5.7471%	2.7027%	4.6512%
Relative absolute error	12.887%	10.9481%	14.695%	12.6846%	17.5041%
Root relative squared error	35.9085%	30.9353%	46.0007%	31.0328%	47.2684%
Total Number of instances	435	435	348	148	43

DIABETES DATASET

J48	10 Cross Validation	Training Set	Percentage Split 20%	Percentage Split 66%	Percentage split 90%
Time Taken	0.06sec	0.03	0.01	0.02	0.02
Number of Leaves	20	20	20	20	20
Size of Tree	39	39	39	39	39
Correctly Classified Instance	567	646	411	199	58
Incorrectly classified Instance	201	122	203	62	19
Correctly classified Percentage	73.8281%	84.1146%	66.9381%	76.2452%	75.3247%
Incorrectly classified Percentage	26.1719%	15.8854%	33.0619%	23.7548%	24.6753%
Relative absolute error	69.4841%	52.4339%	76.3003%	69.2946%	67.737%
Root relative squared error	93.6293%	72.4207%	115.2009%	86.7189%	86.9143%
Total Number of instances	768	768	614	261	77

RANDOM FOREST

IRIS DATASET

Random	10 Cross Validation	Training Set	Percentage Split 20%	Percentage Split 66%	Percentage split 90%
Time Taken	0.09sec	0	0	0	0
Correctly Classified Instance	143	150	116	49	14
Incorrectly classified Instance	7	0	4	2	1
Correctly classified Percentage	95.3333%	100%	96.6667%	96.0784%	93.3333%
Incorrectly classified Percentage	4.6667%	0%	3.3333%	3.9216%	6.6667%
Relative absolute error	9.19%	3.52%	16.7487%	7.8349%	10.9%
Root relative squared error	34.3846%	13.3147%	34.4609%	30.2995%	34.4587%
Total Number of instances	150	150	120	51	15

VOTE DATASET:

Random	10 Cross Validation	Training Set	Percentage Split 20%	Percentage Split 66%	Percentage split 90%
Time Taken	0.11sec	0.1	0.09	0.09	0.14

Correctly Classified Instance	418	432	332	146	40
Incorrectly classified Instance	17	3	16	2	3
Correctly classified Percentage	96.092%	99.3103%	95.4023%	98.6486	93.0233%
Incorrectly classified Percentage	3.908%	0.6897%	4.5977%	1.3514%	6.9767%
Relative absolute error	15.0587%	6.9673%	19.0478%	13.7281%	23.6651%
Root relative squared error	35.7776%	17.733%	39.0502	30.2203%	50.547%
Total Number of instances	435	435	348	148	43

DIABETES DATASET

Random	10 Cross Validation	Training Set	Percentage Split 20%	Percentage Split 66%	Percentage split 90%
Time Taken	0.18sec	0.24	0.23	0.18	0.22
Correctly Classified Instance	582	768	438	205	61
Incorrectly classified Instance	186	0	176	56	16
Correctly classified Percentage	75.7813%	100%	71.3355%	78.5441%	79.2208%
Incorrectly classified Percentage	24.2188%	0%	28.6645%	21.4559%	20.7792%
Relative absolute error	68.3405%	25.0881%	75.2024%	67.5274%	65.2921%
Root relative squared error	84.5604%	31.6064%	89.8177%	82.8626%	82.8001%
Total Number of instances	768	768	614	261	77

In general,machine learning algorithms on datasets with uniform class distribution has high accuracy. Since the algorithm has equal count of instances to learn about each class, its predictions will have more discernment.

Eg. J48 and Random has high accuracy on Iris dataset.

• On other hand, if class distribution is greatly in the favour of a single class, our model will know more about that single class and relatively less about the other classes. Only that class will have desired values in the confusion matrix and all other classes will not have a proper values

• Also, all algorithms are not benefited by equal class spread. Decision Stump split on a single attribute, so a uniform distribution will affect their performance.

Eg. Decision Stump has 66.6% accuracy.

Training and testing methods with cross-validation in focus:

Based on the outcomes of using several training and testing methods available in Weka on the target datasets all this long, it is noticed

10-fold Cross Validation

The original dataset is used for training the classifier using the 10-fold cross validation method.

In this, the dataset is split into 10 sets and for each iteration one set will be test set and other 9 sets will be the training sets. Since this is 10-fold, there will be 10 iterations before the classifier completes the evaluation. Now in each iteration, each set will be used as a test set and other sets are used as training sets. Moreover, this method splits the dataset into subsamples and thus, it ensures a fair class distribution in available in each of the subsets similar to the original dataset.

The actual data set will be divided into specific number of folds. The class distribution will not be maintained as in original data set because during 10 fold cross validation, the datasets are divided and the training and test set differs. This does not affect the classification or the distribution.

Classification Accuracy:

In order to learn about the classification accuracy, different percentage splits are used on different data sets and different classifiers are used. The above data represents the percentage splits and the accuracy determined in each data set.

For Iris data set, the accuracy is more when the percentage split is 90% in Decision stump. In J48 and Random Forest Classifiers the accuracy level is more when the split 66% and in Random Forest it is high when the split is 20%.The error level is also reduced according to the split. Using training set, the accuracy level of the classifier is high.

For Vote data set, the accuracy level is more when the split is nearly 50 or 66% in all the three classifiers.

For Daibetes dataset, the accuracy level is high when the split is 90% in Random Forest and it is high when the percentage split is 66% in Decision stump and J48.

Thus from the analysis the percentage splits affects the classification of the data.

Using Different Parameters:

J48

IRIS DATASET

J48	minNumObj		reducedErrorPruning		doNotMakeSplitPointActualValue	
	2	3	False	True	False	True
Time Taken	0.03 sec	0 sec	0.03 sec	0 sec	0.03 sec	0
Number of Leaves	5	5	5	5	5	5

Size of Tree	9	9	9	9	9	9
Correctly Classified Instance	144	144	144	142	144	145
Incorrectly classified Instance	6	6	6	8	6	5
Correctly classified Percentage	96%	96%	96%	94.6667	96%	96.6667%
Incorrectly classified Percentage	4%	4%	4%	5.3333	4%	3.3333%
Relative absolute error	7.8705%	8.7614%	7.8705%	11.973	7.8705%	6.8705%
Root relative squared error	33.6353%	34.3864%	33.6353%	39.6021	33.6353%	30.5178%
Total Number of instances	150	150	150	150	150	150

minNumObj, reducedErrorPruning and doNotMakeSplitPointActualValue are the important parameters which change the values of the accuracy level.

When minNumObj is increased and reducedErrorPruning is set to true the error percentage increased. When doNotMakeSplitPointActualValue is set to true the accuracy level is increased. The minimum number used for pruning and no.of leaf instance will affect the classifier.

VOTE DATASET

J48	minNumObj		Confidence factor		Unpruned	
	2	4	0.25	0.7	False	True
Time Taken	0.01sec	0 sec	0.01sec	0 sec	0.01sec	0.01
Number of Leaves	6	5	6	10	6	19
Size of Tree	11	9	11	19	11	37
Correctly Classified Instance	419	419	419	418	419	419
Incorrectly classified Instance	16	16	16	17	16	16
Correctly classified Percentage	96.3218	96.3218%	96.3218	96.092	96.3218	96.3218
Incorrectly classified Percentage	3.6782%	3.6782%	3.6782%	3.908	3.6782%	3.6782
Relative absolute error	12.887%	13.3044%	12.887%	12.4223	12.887%	11.6378
Root relative squared error	35.9085%	36.0709%	35.9085%	36.4563	35.9085%	35.8923
Total Number	435	435	435	435	435	435

of instances						

Increase in minNumObj and Confidence factor increased the error level and when unpruned is set to true the error level was decreased to a certain amount. The values used for pruning and leaf will affect the classifiers.

DIABETES DATASET

J48	minNumObj		Confidence factor		Unpruned	
	2	4	0.25	0.5	False	True
Time Taken	0.06sec	0.01sec	0.06sec	0 .02	0.06sec	0
Number of Leaves	20	19	20	22	20	22
Size of Tree	39	37	39	43	39	43
Correctly Classified Instance	567	572	567	559	567	558
Incorrectly classified Instance	201	196	201	209	201	210
Correctly classified Percentage	73.8281%	74.4792%	73.8281%	72.7865	73.8281%	72.6563%
Incorrectly classified Percentage	26.1719%	25.5208%	26.1719%	27.2135	26.1719%	27.3438%
Relative absolute error	69.4841%	69.03%	69.4841%	68.9104	69.4841%	69.2099%
Root relative squared error	93.6293%	91.6243%	93.6293%	94.837	93.6293%	95.3354%
Total Number of instances	768	768	768	768	768	768

In this set, when minNumObj is increased the accuracy level was increased and error was decreased.When the confidence factor was increased, error level was decreased and when unpruned was set to true , error level increased.

Random Forest

Iris Dataset

Random	maxdepth		numIterations		Seed	
	0	2	100	150	1	3
Time Taken	0.09sec	0 .04	0.09sec	0 .04	0.09sec	0.03
Correctly Classified Instance	143	141	143	143	143	142
Incorrectly classified Instance	7	9	7	7	7	8
Correctly classified	95.3333%	94%	95.3333%	95.3333	95.3333%	94.6667%

Percentage						
Incorrectly classified Percentage	4.6667%	6%	4.6667%	4.6667	4.6667%	5.3333%
Relative absolute error	9.19%	12.9114%	9.19%	9.1133	9.19%	9.28%
Root relative squared error	34.3846%	33.173%	34.3846%	33.9962	34.3846%	34.4174%
Total Number of instances	150	150	150	150	150	150

Increase in maxdepth and seed decreased the accuracy and Increase in number of iterations increased the time taken.

VOTE DATASET

Random	maxdepth		numIterations		seed	
	0	2	100	150	1	3
Time Taken	0.11sec	0 .04	0.11sec	0 .13	0.11sec	0.09
Correctly Classified Instance	418	412	418	419	418	419
Incorrectly classified Instance	17	23	17	16	17	16
Correctly classified Percentage	96.092%	94.7126%	96.092%	96.3218%	96.092%	96.3218%
Incorrectly classified Percentage	3.908%	5.2874%	3.908%	3.6782%	3.908%	3.6782%
Relative absolute error	15.0587%	29.0623%	15.0587%	15.0079%	15.0587%	14.6384%
Root relative squared error	35.7776%	42.8566%	35.7776%	35.4619%	35.7776%	35.3293%
Total Number of instances	435	435	435	435	435	435

Increase in max depth decreased the accuracy level and increase in seed increased the accuracy level. Increase in no. of iterations increased the time and accuracy.

DIABETES DATASET

Random	maxdepth		numIterations		seed	
	0	2	100	150	1	3
Time Taken	0.18sec	0 .08	0.18sec	0 .3 sec	0.18sec	0.2
Correctly Classified Instance	582	575	582	579	582	585

Incorrectly classified Instance	186	193	186	189	186	183
Correctly classified Percentage	75.7813%	74.8698%	75.7813%	75.3906%	75.7813%	76.1719%
Incorrectly classified Percentage	24.2188%	25.1302%	24.2188%	24.6094%	24.2188%	23.8281%
Relative absolute error	68.3405%	79.8612%	68.3405%	68.4618%	68.3405%	68.0053%
Root relative squared error	84.5604%	86.0502%	84.5604%	84.5006%	84.5604%	84.2087
Total Number of instances	768	768	768	768	768	768

Increase in maxdepth decreased the accuracy, Increase in iterations decreased the accuracy and increased the time and increase in seed, increased the accuracy.

RANDOM FOREST

NUMBER OF TREES.

IRIS DATA SET

Random		numIterations					
	10	20	50	100	200	500	1000
Time Taken	0.01sec	0 .01	0.01	0.02sec	0 .05	0.1 sec	0.18
Correctly Classified Instance	143	143	142	143	143	143	143
Incorrectly classified Instance	7	7	8	7	7	7	7
Correctly classified Percentage	95.3333%	95.3333%	94.6667%	95.3333%	95.3333%	95.3333%	95.3333%
Incorrectly classified Percentage	4.6667%	4.6667%	5.3333%	4.6667%	4.6667%	4.6667%	4.6667%
Relative absolute error	9.4%	9.15%	9.18%	9.19%	9.02%	8.952%	8.997%
Root relative squared error	35.0999%	35.3058%	34.909%	34.3846%	33.8012%	33.6841%	33.7219%
Total Number of instances	150	150	150	150	150	150	150

The increase in iterations, increased the time taken, reduced the error level and increased the accuracy.

VOTE DATASET

Random		numIterations					
	10	20	50	100	200	500	1000
Time Taken	0.03sec	0 .03	0.06	0.09	0 .12	0.3 sec	0.57
Correctly Classified Instance	419	419	420	418	420	419	420
Incorrectly classified Instance	16	16	15	17	15	16	15
Correctly classified Percentage	96.3218	96.3218%	96.5517%	96.092%	96.5517%	96.3218	96.5517%
Incorrectly classified Percentage	3.6782%	3.6782%	3.4483%	3.908%	3.4483%	3.6782	3.4483%
Relative absolute error	13.75%	14.3916%	14.6637%	15.0587%	15.1608%	15.191%	15.099%
Root relative squared error	35.7374%	35.6245%	35.229%	35.7776%	35.664%	35.9526%	35.941%
Total Number of instances	435	435	435	435	435	435	435

The increase in iterations, increased the time taken, reduced the error level and increased the accuracy.

DIABETES DATASET

Random		numIterations					
	10	20	50	100	200	500	1000
Time Taken	0.05sec	0 .06	0.11	0.23	0 .35	0.92sec	1.85sec
Correctly Classified Instance	571	575	582	582	581	582	579
Incorrectly classified Instance	197	193	186	186	187	186	189
Correctly classified Percentage	74.349	74.8698%	75.7813%	75.7813%	75.651%	75.7813%	75.3906%
Incorrectly classified Percentage	25.651%	25.1302%	24.2188%	24.2188%	24.349%	24.2188%	24.6094%
Relative absolute error	68.584%	68.5411%	68.5095%	68.3405%	68.4494%	68.5783%	68.6161%
Root relative squared error	88.4941%	86.1801%	85.1073%	84.5604%	84.3954%	84.5028%	84.5392%
Total	768	768	768	768	435	768	768

Number of instances							

The increase in iterations, increased the time taken, reduced the accuracy level and increased the error.

More number of iterations will be good for more accuracy.

Max Depth:

IRIS DATASET

Random	Depth 0	2	4	0	2	4	0	2	4
	Iteration 10	10	10	100	100	100	500	500	500
Time Taken	0.01sec	0 .01	0.	0.02	0.01	0.03	0.05	0.09	0.1
Correctly classified Percentage	95.3333%	94%	96%	95.33	94	95.33	95.33	94	95.33
Incorrectly classified Percentage	4.6667%	6%	4%	4.66	6	4.66	4.66	6	4.66
Relative absolute error	9.4%	12.6582%	9.2801%	9.19	12.91	9.16	8.952	12.89	9.03
Root relative squared error	35.0999%	33.4679%	34.18%	34.38	33.173	33.51	33.68	32.95	33.16

VOTE DATASET

Random	Depth 0	2	4	0	2	4	0	2	4
	Iteration 10	10	10	100	100	100	500	500	500
Time Taken	0.03	0 .01	0.01	0.08	0.03	0.05	0.32	0.07	0.13
Correctly classified Percentage	96.32	95.63	95.63	96.09	94.71	96.09	96.32	94.94	95.86
Incorrectly classified Percentage	3.67	4.36	4.3	3.90	5.28	3.90	3.67	5.05	4.13
Relative absolute error	13.75	26.41	18.86	15.05	29.06	19.28	15.191	29.088	19.32
Root relative squared error	35.73	41.58	36.73	35.77	42.85	36.53	35.95	42.7941	36.755

DIABETES DATASET

Random	Depth 0	2	4	0	2	4	0	2	4
	Iteration 10	10	10	100	100	100	500	500	500
Time Taken	0.02	0 .22	1.03	0.01	0.06	0.32	0.03	0.12	0.54
Correctly classified Percentage	74.34	75.78	75.78	75.13	74.86	75.78	76.17	75.65	76.30
Incorrectly classified Percentage	25.65	24.21	424.21	24.86	25.13	24.21	23.82	24.34	23.69
Relative absolute error	68.58	68.34	68.57	79.78	79.86	79.81	71.05	71.72	71.99
Root relative squared error	88.49	84.56	84.50	87.08	86.05	85.93	83.98	83.18	83.41

From all the three dataset it is found that maxdepth should be minimum and the number of trees should be more for the accuracy level to be high. The maxdepth and the number of trees will work together if depth is set to minimum and number of trees is increased.The are indirectly proportional to each other.

4. Noise and Missing Values

Apart from the original datasets, two more versions of the same are used with the decision tree algorithms. One variant has missing values and the other has considerable noise (some instances are misclassified). Three testing options are used and the test results for all three dataset using Decision Stump, J48 and Random Forest classifiers are tabulated below.

5%, 10% and 50 missing values were introduced in petallength attribute.

Decision Stump – Missing Values

Classification	Normal	5%Missing	10% Missing	50% Missing
	petallength <= 2.45: Iris-setosa petallength <= 2.45: Iris-setosa petallength is missing: Iris-setosa	petallength <= 2.45: Iris-setosa petallength >2.45: Iris-versicolor petallength is missing: Iris-setosa	petallength <= 2.45: Iris-setosa petallength >2.45: Iris-versicolor petallength is missing: Iris-setosa	petalwidth <= 4.75: Iris-setosa petalwidth >4.75: Iris-versicolor petalwidth is missing: Iris-setosa

Class Distribution:

Normal

Class Distribution	Iris-setosa	Iris-versicolor	Iris-virginica
petallength <= 2.45	1.0	0.0	0.0
petallength > 2.45	0.0	0.5	0.5
petallength is missing	0.333333	0.3333333	0.333333

5% Missing Data

Class Distribution	Iris-setosa	Iris-versicolor	Iris-virginica
petallength <= 2.45	1.0	0.0	0.0
petallength > 2.45	0.0	0.5	0.5
Petallength is missing	1.0	0.0	0.0

10% Missing Data

Class Distribution	Iris-setosa	Iris-versicolor	Iris-virginica
petallength <= 2.45	1.0	0.0	0.0
petallength > 2.45	0.0	0.5	0.5
petallength ismissing	1.0	0.0	0.0

50% Missing Data

Class Distribution	Iris-setosa	Iris-versicolor	Iris-virginica
petallength <= 4.75	0.0	0.9565217391304348	0.043478260869565216
petallength > 4.75	0.0	0.057692307692307696	0.9423076923076923
petallength is missing	0.6666666666666666	0.33333333333333333	0.0

When there is no missing data, 5% missing data and 10% missing data the summary is as below:

Summary

Correctly Classified Instances	100	66.6667%
Incorrectly Classified Instances	50	33.3333%
Kappa statistic	0.5	
Mean absolute error	0.2222	
Root mean squared error	0.3333	
Relative absolute error		50%
Root relative squared error		70.7107%

Confusion Matrix

Classified as	A	b	C
a = Iris-setosa	50	0	0
b = Iris-versicolor	0	50	0
c = Iris-virginica	0	50	0

When 50 % of data is missing

Summary

Correctly Classified Instances	118	78.6667 %
Incorrectly Classified Instances	32	21.3333 %
Kappa statistic	0.68	
Mean absolute error	0.1946	
Root mean squared error	0.3216	
Relative absolute error		43.7816 %
Root relative squared error		68.2156 %

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	50	0	0
b = Iris-versicolor	24	21	4
c = Iris-virginica	0	3	47

When decision stump is used for classifying, the accuracy is less. When more amount of missing values are introduced, classifier selects the appropriate attribute and classifies it. From above classification the Accuracy is more and error is reduced when 50% of missing data is introduced.

J48 Missing Values

Missing Values are introduced to petallength attribute.

When there is no data missing, 5% data missing and 10% data missing

J48 pruned tree

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
petalwidth <= 1.7
petallength <= 4.9: Iris-versicolor (48.0/1.0)
petallength > 4.9
petalwidth <= 1.5: Iris-virginica (3.0)
petalwidth > 1.5: Iris-versicolor (3.0/1.0)
petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves	5	
Size of the tree	9	
Correctly Classified Instances	144	96%
Incorrectly Classified Instances	6	4%
Kappa statistic	0.94	
Mean absolute error	0.035	
Root mean squared error	0.1586	
Relative absolute error		7.8705%
Root relative squared error		33.6353%

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	49	1	0
b = Iris-versicolor	0	47	3
c = Iris-virginica	0	2	48

For 50% Missing data

J48 Pruned tree

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
petalwidth <= 1.7: Iris-versicolor (54.0/5.0)
petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves:	3	
Size of the tree:	5	
Correctly Classified Instances	142	94.6667 %
Incorrectly Classified Instances	8	5.3333 %
Kappa statistic	0.92	
Mean absolute error	0.0576	
Root mean squared error	0.1832	
Relative absolute error	12.9526 %	
Root relative squared error	38.8576 %	
Total Number of Instances	150	

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	49	1	0
b = Iris-versicolor	0	48	2
c = Iris-virginica	0	5	45

J48 classifier classifies the data accurately when compared to decision stump classifier. When missing values were introduced to petallength attribute, accuracy level was almost same and the error percentage was less. Number of leaves and size of the tree was also more. When 50% of missing data was introduce, the accuracy level dropped by 2% and the error level increased by 5%. Number of leaves and size of the tree was also reduced. Thus when more missing values are introduced the accuracy of classification is reduced.

Random Forest

Missing Values are introduced to petallength attribute.

Summary

	No missing data		5% missing data		10% missing data		50% missing data	
Correctly Classified Instances	143	95.3333%	143	95.3333%	143	95.3333%	142	94.6667 %
Incorrectly	7	4.6667%	7	4.6667%	7	4.6667%	8	5.3333 %

Classified Instances								
Time taken to build model	0.13s		0.06s		0.06s		0.04s	
Mean absolute error	0.0408		0.0422		0.0438		0.0532	
Root mean squared error	0.1621		0.1612		0.1631		0.1788	
Relative absolute error		9.19%		9.4889%		9.8515%		11.9798%
Root relative squared error		34.3846%		34.193%		34.5892%		37.9397%

No data missing, 5% data, 10% data

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	50	0	0
b = Iris-versicolor	0	47	3
c = Iris-virginica	0	4	46

When 50% data is missing

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	50	0	0
b = Iris-versicolor	0	47	3
c = Iris-virginica	0	5	45

When missing values are introduced to petallength attribute the accuracy level of the classifier was reduced and the error percentage increased gradually. When 50% of missing values were introduced, the accuracy level reduced by 1% and the error level increased by 3 %.

So, it is noticed for all the classifiers when certain amount of missing values are introduced, the accuracy of the classifier drops and the error percentage is increased.

Introduce Noise

The noise was introduce to the Iris dataset by misclassifying some datas.5%, 10% and 50% of noise was introduced and through classification the following data were collected.

Decision Stump

Classification	Normal	5% Noise	10% Noise	50% Noise

	petallength <= 2.45: Iris-setosa petallength <= 2.45: Iris-setosa petallength is missing: Iris-setosa	petallength <= 4.75: Iris-versicolor petallength >4.75: Iris-virginica petallength is missing: Iris-versicolor	petallength <= 4.75: Iris-versicolor petallength >4.75: Iris-virginica petallength is missing: Iris-versicolor	petallength <= 4.85: Iris-versicolor petallength >4.85: Iris-virginica petallength is missing: Iris-virginica

Summary – No Noise

Time taken to build model	0s	
Correctly Classified Instances	100	66.6667 %
Incorrectly Classified Instances	50	33.3333 %
Kappa statistic	0.5	
Mean absolute error	0.2222	
Root mean squared error	0.3333	
Relative absolute error		50%
Root relative squared error		70.7107%

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	50	0	0
b = Iris-versicolor	0	50	0
c = Iris-virginica	0	50	0

5% Noise Summary

Time taken to build model	0s	
Correctly Classified Instances	98	65.3333%
Incorrectly Classified Instances	52	34.6667%
Kappa statistic	0.4513	
Mean absolute error	0.2705	
Root mean squared error	0.3765	
Relative absolute error		61.3566%
Root relative squared error		80.19%

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	0	42	0
b = Iris-versicolor	0	52	6
c = Iris-virginica	0	4	46

10% Noise

Time taken to build model	0.01s	
Correctly Classified Instances	104	69.3333%
Incorrectly Classified Instances	46	30.6667%
Kappa statistic	0.4918	
Mean absolute error	0.2639	

Root mean squared error	0.3706	
Relative absolute error		61.0299%
Root relative squared error		79.7082%
Total Number of Instances	150	

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	0	35	0
b = Iris-versicolor	0	58	6
c = Iris-virginica	0	5	46

50% Noise

Time taken to build model	0s	
Correctly Classified Instances	91	60.6667%
Incorrectly Classified Instances	59	39.3333%
Kappa statistic	0.354	
Mean absolute error	0.3319	
Root mean squared error	0.4111	
Relative absolute error		81.8651%
Root relative squared error		91.3769%

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	0	17	7
b = Iris-versicolor	0	49	1
c = Iris-virginica	0	34	42

The noise was introduced and the decision stump classifier was used. In this, it is found when 5% and 10% of noise was introduced the accuracy level increased and the error percentage was reduced. But when 50% of noise was introduced the accuracy level dropped by 6% and the error level was increased by 25%. Thus making it difficult for further analysis.

J48

J48 pruned tree

No noise

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
petalwidth <= 1.7
petallength <= 4.9: Iris-versicolor (48.0/1.0)
petallength > 4.9
petalwidth <= 1.5: Iris-virginica (3.0)
petalwidth > 1.5: Iris-versicolor (3.0/1.0)
petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves	5	

Size of the tree	9	
Correctly Classified Instances	144	96%
Incorrectly Classified Instances	6	4%
Kappa statistic	0.94	
Mean absolute error	0.035	
Root mean squared error	0.1586	
Relative absolute error		7.8705%
Root relative squared error		33.6353%
Total Number of Instances	150	

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	50	0	0
b = Iris-versicolor	0	50	0
c = Iris-virginica	0	50	0

5% Noise

Number of Leaves	5	
Size of the tree	9	
Correctly Classified Instances	133	88.6667%
Incorrectly Classified Instances	17	11.3333%
Kappa statistic	0.8299	
Mean absolute error	0.1099	
Root mean squared error	0.2641	
Relative absolute error		24.94%
Root relative squared error		56.249%

Time taken to build model: 0.02s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	40	2	0
b = Iris-versicolor	8	45	5
c = Iris-virginica	0	2	48

10% Noise

Number of Leaves	5	
Size of the tree	9	
Correctly Classified Instances	126	84%
Incorrectly Classified Instances	24	16%
Kappa statistic	0.7583	
Mean absolute error	0.1394	
Root mean squared error	0.2925	
Relative absolute error		32.2332%
Root relative squared error		62.9109%

Time taken to build model: 0.01s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	31	4	0
b = Iris-versicolor	14	47	3
c = Iris-virginica	1	2	48

50% Noise

petallength <= 4.8
petalwidth <= 0.6
sepallength <= 5.6: Iris-virginica (47.0/17.0)
sepallength > 5.6: Iris-versicolor (3.0)
petalwidth > 0.6
petalwidth <= 1.6: Iris-versicolor (45.0/16.0)
petalwidth > 1.6: Iris-virginica (4.0/1.0)
petallength > 4.8
petalwidth <= 1.7: Iris-setosa (8.0/3.0)
petalwidth > 1.7: Iris-virginica (43.0/3.0)

Number of Leaves	6	
Size of the tree	11	
Correctly Classified Instances	94	62.6667%
Incorrectly Classified Instances	56	37.3333%
Kappa statistic	0.3406	
Mean absolute error	0.2859	
Root mean squared error	0.4094	
Relative absolute error		70.5194%
Root relative squared error		91.0152%

Time taken to build model: 0s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	1	16	7
b = Iris-versicolor	2	27	21
c = Iris-virginica	4	6	66

When noise was introduced in data and J48 is used to classify the data,it is found that the accuracy level started dropping and the error level was increased gradually. So when more noise (50%)was introduced the accuracy level dropped by 35% and the error percentage increased by 40 to 50%. Thus it is clearly understood that when more noise is introduced in the dataset, it makes the dataset unfit for classification.

Random Forest:

No Noise

Correctly Classified Instances	143	95.3333%
Incorrectly Classified Instances	7	4.6667%

Kappa statistic	0.13s	
Mean absolute error	0.0408	
Root mean squared error	0.1621	
Relative absolute error		9.19%
Root relative squared error		34.3846%

5% Noise

Correctly Classified Instances	131	87.3333%
Incorrectly Classified Instances	19	12.6667%
Kappa statistic	0.8092	
Mean absolute error	0.1152	
Root mean squared error	0.2687	
Relative absolute error		26.1422%
Root relative squared error		57.2311%

Time taken to build model: 0.04 s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	39	3	0
b = Iris-versicolor	8	47	3
c = Iris-virginica	0	5	45

10% Noise

Correctly Classified Instances	120	80%
Incorrectly Classified Instances	30	20%
Kappa statistic	0.6926	
Mean absolute error	0.1474	
Root mean squared error	0.2992	
Relative absolute error		34.094%
Root relative squared error		64.3585%

Time taken to build model: 0.06 s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	27	8	0
b = Iris-versicolor	10	49	5
c = Iris-virginica	1	6	44

50% Noise

Correctly Classified Instances	88	58.6667%
Incorrectly Classified Instances	62	41.3333%
Kappa statistic	0.3147	
Mean absolute error	0.2931	
Root mean squared error	0.4335	
Relative absolute error		72.2997%
Root relative squared error		96.3736%

Time taken to build model: 0.05 s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	5	14	5
b = Iris-versicolor	7	30	13
c = Iris-virginica	4	19	53

When noise was introduced in data and Random Forest is used to classify the data, it is found that the accuracy level started dropping and the error level was increased gradually. So when more noise (50%)was introduced the accuracy level dropped by 30% and the error percentage increased by 45%. Thus it is clearly understood that when more noise is introduced in the dataset, it makes the dataset unfit for classification.

From the whole analysis though decision stump does not classify accurately, when more noise was introduced the accuracy level was reduced. In J48 and Random forest when noise is introduced accuracy level decreased from starting.

Range of an attribute increased to 1000 times.

With 50% noise introduced, the petallength values are increased 1000 times more than the previous value. So the range of the of the petallength is in a different scale than the others.

Decision Stump

Classifications
petallength <= 4850.0: Iris-versicolor
petallength >4850.0: Iris-virginica
petallength is missing: Iris-virginica

Correctly Classified Instances	91	60.6667%
Incorrectly Classified Instances	59	39.3333%
Kappa statistic	0.354	
Mean absolute error	0.3319	
Root mean squared error	0.4111	
Relative absolute error		81.8651%
Root relative squared error		91.3769%

Time taken to build model: 0.01s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	0	17	7
b = Iris-versicolor	0	49	1
c = Iris-virginica	0	34	42

J48

petallength <= 4800
petalwidth <= 0.6
sepallength <= 5.6: Iris-virginica (47.0/17.0)
sepallength > 5.6: Iris-versicolor (3.0)
petalwidth > 0.6
petalwidth <= 1.6: Iris-versicolor (45.0/16.0)
petalwidth > 1.6: Iris-virginica (4.0/1.0)
petallength > 4800
petalwidth <= 1.7: Iris-setosa (8.0/3.0)
petalwidth > 1.7: Iris-virginica (43.0/3.0)

Number of Leaves	6	
Size of the tree	11	
Correctly Classified Instances	94	62.6667%
Incorrectly Classified Instances	56	37.3333%
Kappa statistic	0.3406	
Mean absolute error	0.2859	
Root mean squared error	0.4094	
Relative absolute error		70.5194%
Root relative squared error		91.0152%

Time taken to build model: 0.01s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	1	16	7
b = Iris-versicolor	2	27	21
c = Iris-virginica	4	6	66

Random Forest

Correctly Classified Instances	87	58%
Incorrectly Classified Instances	63	42%
Kappa statistic	0.2996	
Mean absolute error	0.2972	
Root mean squared error	0.4365	
Relative absolute error		73.3018%
Root relative squared error		97.0266%

Time taken to build model: 0.06s

Confusion Matrix

Classified as	A	B	C
a = Iris-setosa	4	15	5
b = Iris-versicolor	7	29	14
c = Iris-virginica	4	18	54

By comparing the values, 50% noise data classification using decision stump is almost equal to the one 50% noise introduced data classification in which petallength is increased to 1000 times. Even J48 and Random Forest classifications are also same except wwith the difference of 1% or 2 %.

This is because when one attribute range is different from the range of other attributes, we cannot classify it using the range. It may lead to wrong classification. In order to avoid that, the classifiers use normalisation. By using Normalisation, the variance and standard deviations are used and it is used by the classifier. Because of that, though the range of an attribute is increased 1000 times, it still gives the same accuracy when classified using the decision trees.

3. FEATURE SELECTION

InfoGainAttributeEval : Evaluates the worth of an attribute by measuring the information gain with respect to the class.

CorrelationAttributeEval : Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

By measuring the information gain and correlation of the attributes, the attributes will be selected for the classifier. The three data sets are used and the attribute are selected through attribute evaluation.

IRIS DATASET

Correlation Attribute Eval

Attribute Evaluator (supervised, Class (nominal): 5 class):

Correlation Ranking Filter

Ranked	Attributes
0.615	3 petallength
0.592	4 petalwidth
0.478	1 sepallength
0.397	2 sepalwidth

Selected attributes: 3,4,1,2: 4

Infogain Attribute Eval

Attribute Evaluator (supervised, Class (nominal): 5 class):

Information Gain Ranking Filter

Ranked	Attributes
1.418	3 petallength
1.378	4 petalwidth
0.698	1 sepallength
0.376	2 sepalwidth

Selected attributes: 3,4,1,2: 4

VOTE DATASET

Correlation Attribute Eval

Attribute Evaluator (supervised, Class (nominal): 17 Class):

Correlation Ranking Filter

Ranked	Attributes
0.9096	4 physician-fee-freeze
0.7343	3 adoption-of-the-budget-resolution
0.6837	5 el-salvador-aid
0.6666	12 education-spending
0.6283	9 mx-missile
0.617	8 aid-to-nicaraguan-contras
0.6063	14 crime
0.5268	13 superfund-right-to-sue
0.5127	15 duty-free-exports
0.5045	7 anti-satellite-test-ban
0.413	6 religious-groups-in-schools
0.3931	1 handicapped-infants
0.3669	11 synfuels-corporation-cutback
0.3519	16 export-administration-act-south-africa
0.0838	10 immigration
0.011	2 water-project-cost-sharing

Selected attributes: 4,3,5,12,9,8,14,13,15,7,6,1,11,16,10,2: 16

Infogain Attribute Eval:

Attribute Evaluator (supervised, Class (nominal): 17 Class):

Information Gain Ranking Filter

Ranked	Attributes
0.7078541	4 physician-fee-freeze
0.4185726	3 adoption-of-the-budget-resolution
0.4028397	5 el-salvador-aid
0.34036	12 education-spending
0.3123121	14 crime
0.3095576	8 aid-to-nicaraguan-contras
0.2856444	9 mx-missile
0.2121705	13 superfund-right-to-sue
0.2013666	15 duty-free-exports
0.1902427	7 anti-satellite-test-ban
0.1404643	6 religious-groups-in-schools
0.1211834	1 handicapped-infants
0.1007458	11 synfuels-corporation-cutback
0.0529956	16 export-administration-act-south-africa
0.0049097	10 immigration

0.0000117	2 water-project-cost-sharing

Selected attributes: 4,3,5,12,14,8,9,13,15,7,6,1,11,16,10,2: 16

DIABETES DATASET

Correlation Attribute Eval

Attribute Evaluator (supervised, Class (nominal): 9 class):

Correlation Ranking Filter

Ranked	Attribute
0.4666	2 plas
0.2927	6 mass
0.2384	8 age
0.2219	1 preg
0.1738	7 pedi
0.1305	5 insu
0.0748	4 skin
0.0651	3 pres

Selected attributes: 2,6,8,1,7,5,4,3: 8

InfoGain Attribute Eval

Attribute Evaluator (supervised, Class (nominal): 9 class):

Information Gain Ranking Filter

Ranked	Attribute
0.1901	2 plas
0.0749	6 mass
0.0725	8 age
0.0595	5 insu
0.0443	4 skin
0.0392	1 preg
0.0208	7 pedi
0.014	3 pres

Selected attributes: 2,6,8,5,4,1,7,3: 8

The Correlation Attribute Eval and InfoGain Attribute Eval gives the result in terms of ranks. Using this ranking factor, the attributes are selected. The attribute with highest rank has to be selected by the classifier, inorder to classify the data accurately. If the attributes with lower ranks are selected then the accuracy of the classification will be reduced.

By Analysying and Visualizing data it is noticed that in each classifier the selected attributes are used to classify the data sets which increases the Classification accuracy and decreases the error percentage.