Machine Learning Assignment 1

Steps:

1. Discretization:
    1. For each attribute in dataset, minimum and maximum value of attribute is calculated.
    2. Difference in minimum and maximum value is calculated. diff=(max-min)
    3. K intervals in data are generated using *interval=diff/k*.
    4. Data points are classified into the interval to which they belong.

2. Randomization:
    1. For getting a random set of datasets we have first used the inbuilt function *randperm()* on each category of flowers and then divided the set into two sets of 25 each namely Training and Test.
    2. Repeat above 100 times to generate 100 training and 100 test data for each type of flower.

3. Find-S Algorithm:
    1. Generate a hypothesis for each training data.
    2. Test the hypothesis using the corresponding test data.

Results:

For a random experiment it was observed that:

Value of K	Setosa Accuracy			Versicolor Accuracy			Virginica Accuracy		
	min	max	avg	min	max	avg	min	max	Avg
3	24	25	24.95	25	25	25	24	25	24.91
5	23	25	24.91	25	25	25	24	25	24.90
7	23	25	24.94	24	25	24.98	25	25	25
9	23	25	24.96	24	25	24.97	25	25	25

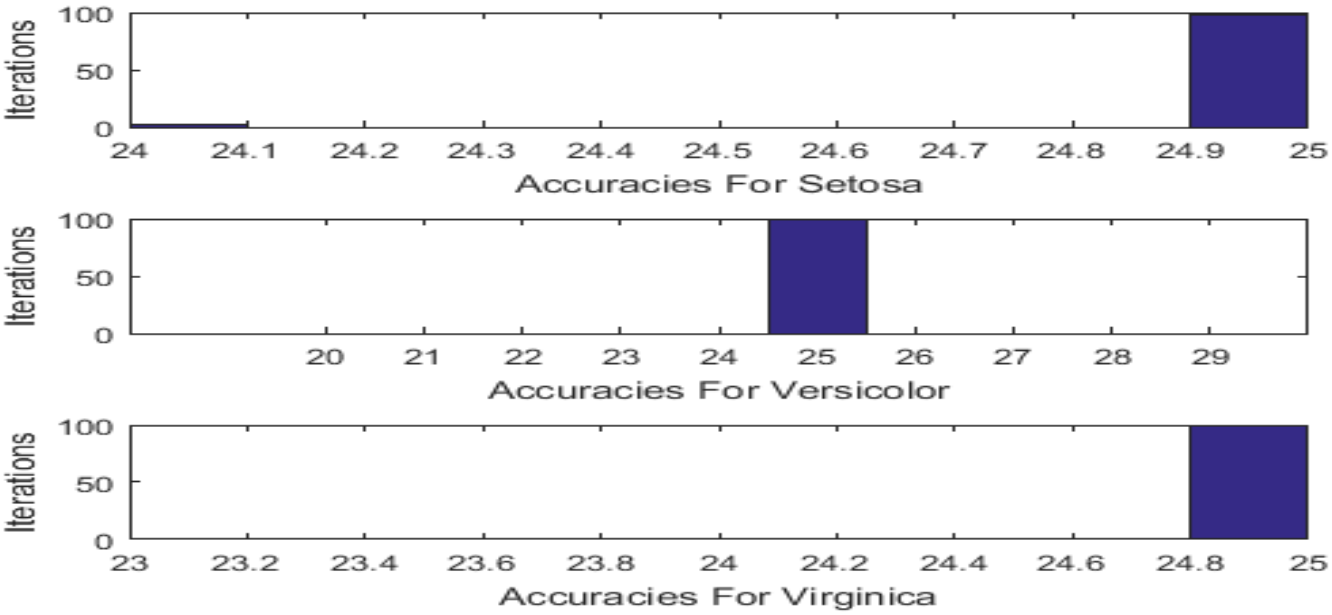
- The value of k has little effect on the accuracy of the algorithm. Distribution of the data entries in the dataset are equally responsible for determining the accuracy. A more generic hypothesis gives a more accurate result compared to a more specific hypothesis.
- Among all the three types of flowers in iris dataset, “iris-versicolor” shows maximum accuracy for all values of k. This is because the data points in test dataset are very close to the hypothesis generated using training dataset. For any value of k, the hypothesis generated has data points from each interval, and this makes the hypothesis more general than being specific. For ex: iris-versicolor has a hypothesis of type:

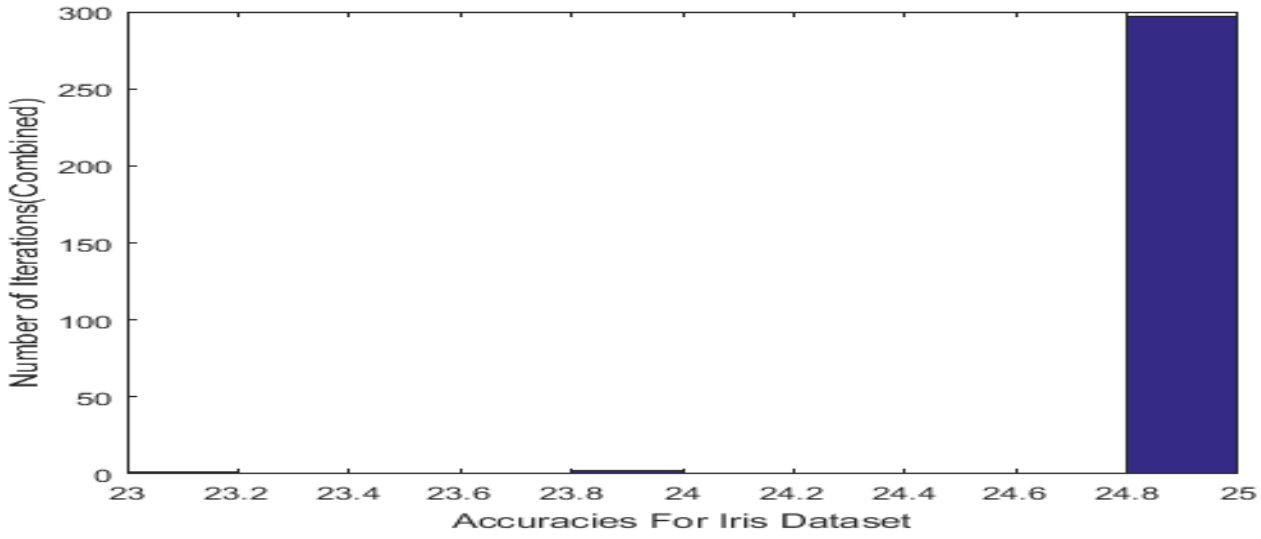
  h=[{1,2,3},{1,2},{1,2,3},{1,2,3}] for k=3
  h=[{1,2,3,4,5},{1,2,4,5},{1,2,3,4,5},{1,2,3,5}] for k=5

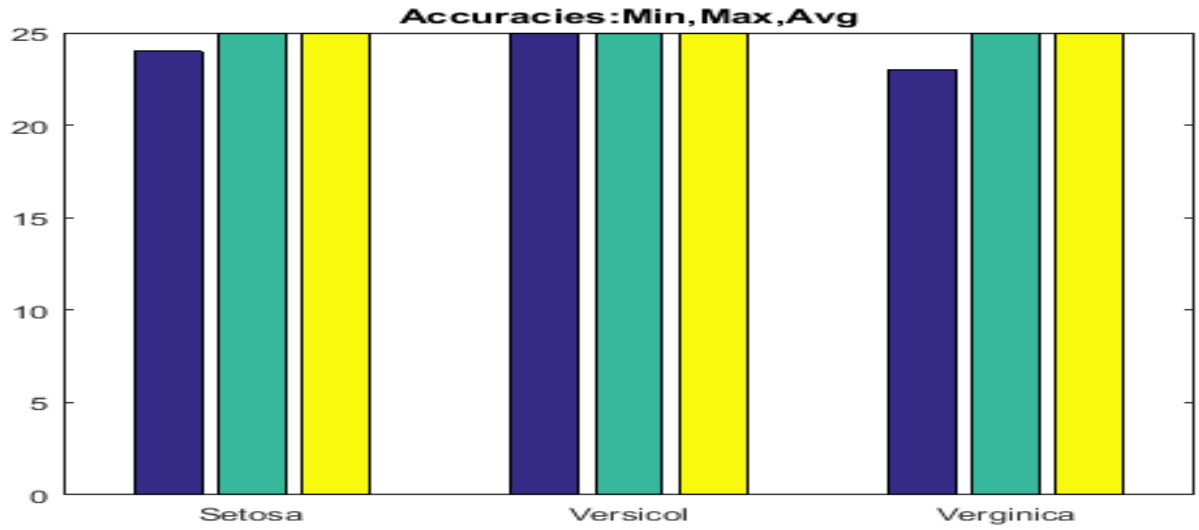
  This hypothesis is more close to general than specific. In this case, a new test dataset is more likely to satisfy the hypothesis as the hypothesis supports all types of data points.

```
load iris.dat % Loading Data Set
newDat = iris; % Variable To store Discretized data
for k=[3,5,7,9] % Iterating for each value of K
    for i = 1:4 % Discretization Step
        mini = min(newDat(:,i));
        maxi = max(newDat(:,i));
        diff = (maxi-mini)/k;
        for j = 1:length(newDat(:,i))
            t=0;
            for p = 1:k
                if newDat(j,i)<=(p*diff+mini) && newDat(j,i)>=((p-1)*diff+mini) % Condition to check where does each data point fits in Intervals
                    t=p;
                end
            end
            newDat(j,i)=t;
        end
    end
    accuracy = {[],[],[]}; % Variable to store accuracies for each of the flower type
    for i=1:3 % Iterating over each Flower type
        hypo = {[],[],[],[]}; % Initial Hypothesis
        flower = newDat((1+50*(i-1)):50*i,:);
        for x=1:100 % Find – S Algorithm executed 100 times for each of the Flower type
            randFlower = flower(randperm(size(flower,1)),:);
            trainFlower = randFlower(1:25,:);
            testFlower = randFlower(26:end,:);
            accuracyCount=0;
            % Training of Algorithm to generate Hypothesis as Output based on Training Data
            for j=1:length(trainFlower)
                if ismember(trainFlower(j,1),hypo{1})==0
                    hypo{1}(end+1) = trainFlower(j,1);
                end
                if ismember(trainFlower(j,2),hypo{2})==0
                    hypo{2}(end+1) = trainFlower(j,2);
                end
                if ismember(trainFlower(j,3),hypo{3})==0
                    hypo{3}(end+1) = trainFlower(j,3);
                end
                if ismember(trainFlower(j,4),hypo{4})==0
                    hypo{4}(end+1) = trainFlower(j,4);
                end
            end
            % Testing of generated Hypothesis based on Test Data
            for j=1:length(testFlower)
                % Condition to check whether the test data modifies the hypothesis, if it modifies, it is considered as Negetive Data
                if ((ismember(testFlower(j,1),hypo{1})~=0) && (ismember(testFlower(j,2),hypo{2})~=0) && (ismember(testFlower(j,3),hypo{3})~=0) &&
(ismember(testFlower(j,4),hypo{4})~=0))
                    accuracyCount=accuracyCount+1 ;
                end
            end
            accuracy{i}(end+1)=accuracyCount; % Appending Accuracy Count for each iteration for each of the Flower Type
        end
    end
    combined = horzcat(accuracy{1},accuracy{2},accuracy{3}); % Combining Accuracy for Combined result
    minSetosa = min(accuracy{1});
    minVersicolor = min(accuracy{2});
    minVerginica = min(accuracy{3});

    maxSetosa = max(accuracy{1});
    maxVersicolor = max(accuracy{2});
    maxVerginica = max(accuracy{3});

    avgSetosa = mean(accuracy{1});
    avgVersicolor = mean(accuracy{2});
    avgVerginica = mean(accuracy{3});

    figure() % Figure to Compare accuracies of each Flower type
    Labels = {'Setosa','Versicolor','Virginica'};
    b = bar([minSetosa maxSetosa avgSetosa; minVersicolor maxVersicolor avgVersicolor; minVerginica maxVerginica avgVerginica]);
    set(gca,'XTickLabel', Labels);
    ax = gca;
    ax.Title.String = 'Accuracies:Min,Max,Avg';

    figure() % Figure to Show accuracy of Individual Flower type
    hist(combined)
    xlabel('Accuracies For Iris Dataset')
    ylabel('Number of Iterations(Combined)')

    figure()
    subplot(3,1,1)
    hist(accuracy{1})
    xlabel('Accuracies For Setosa')
    ylabel('Iterations')
    subplot(3,1,2)
    hist(accuracy{2})
    xlabel('Accuracies For Versicolor')
    ylabel('Iterations')
    subplot(3,1,3)
    hist(accuracy{3})
    xlabel('Accuracies For Virginica')
    ylabel('Iterations')
end
```

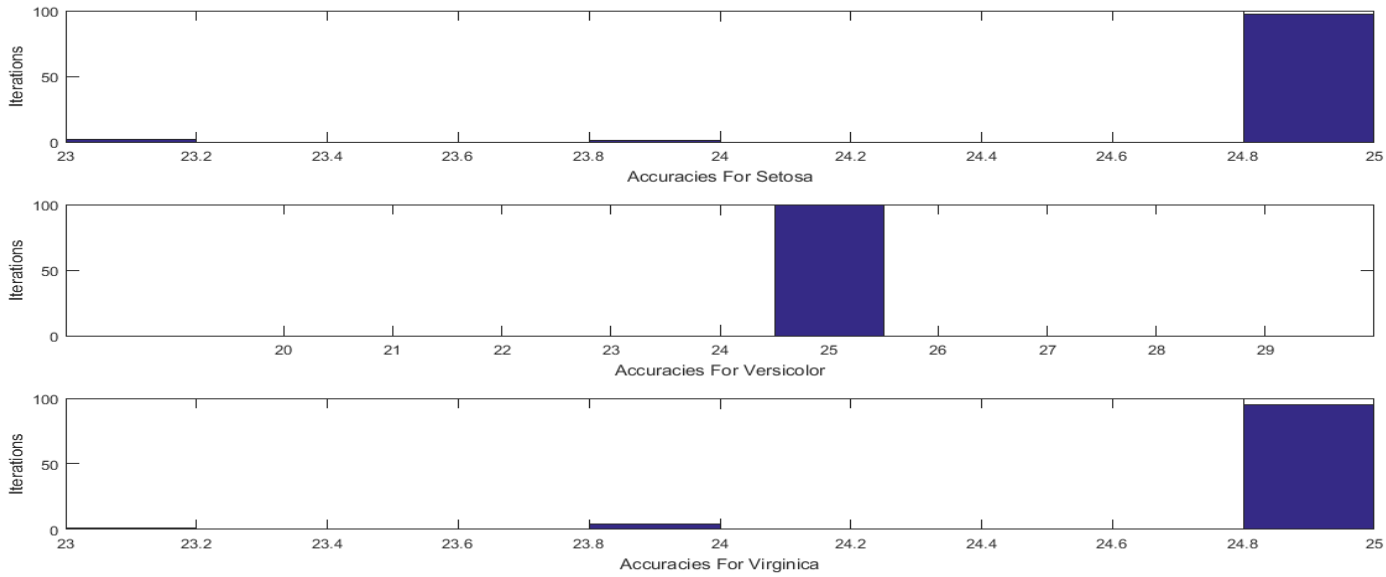
For K = 3:

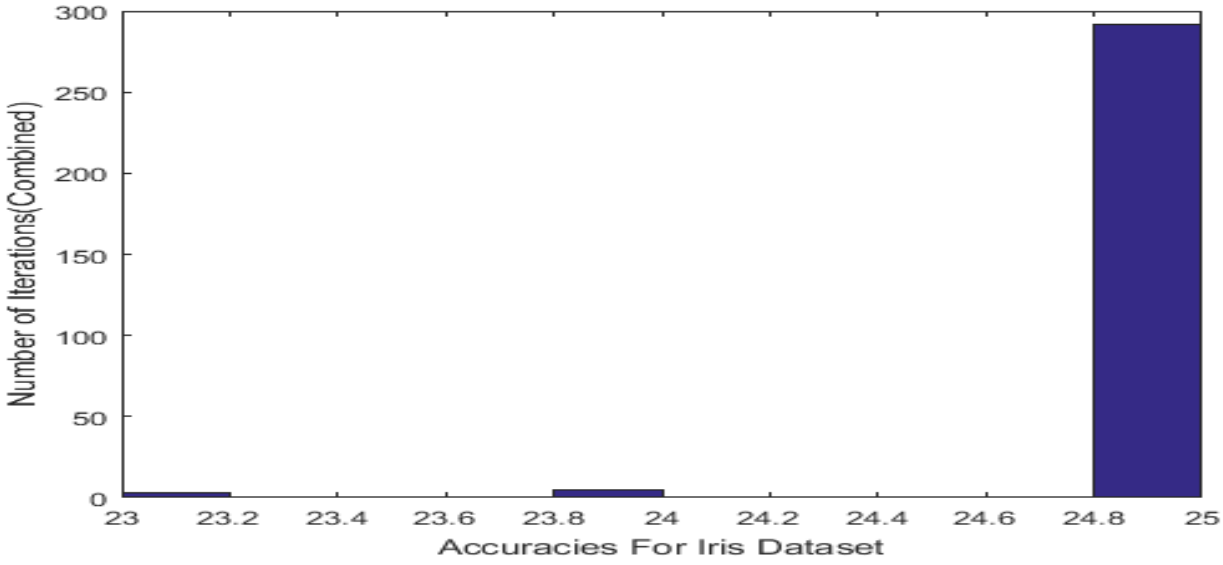

Iterations
Accuracies For Setosa
Iterations
Accuracies For Versicolor
Iterations
Accuracies For Virginica

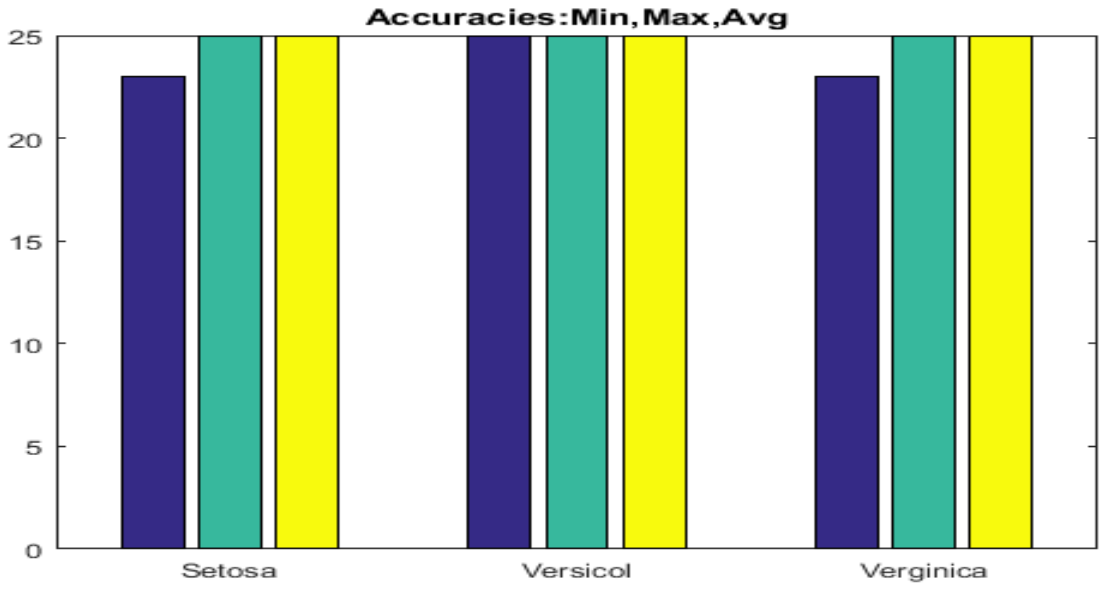

Number of Iterations(Combined)
Accuracies For Iris Dataset


Accuracies:Min,Max,Avg
Setosa
Versicol
Verginica

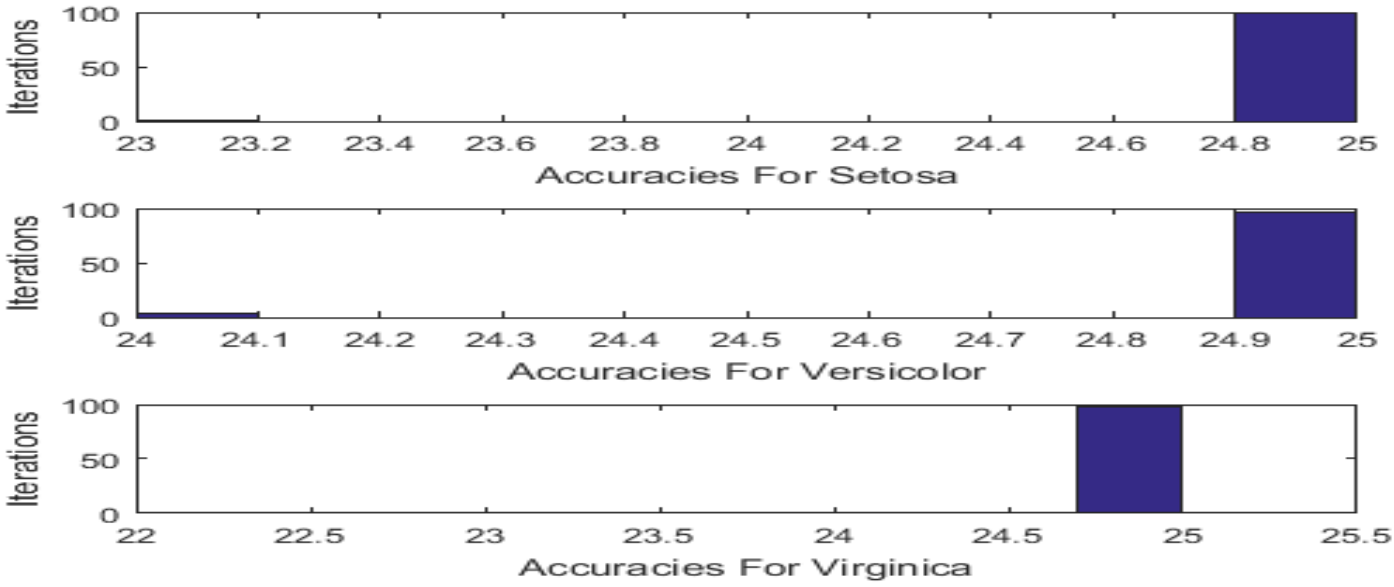
For K = 5:

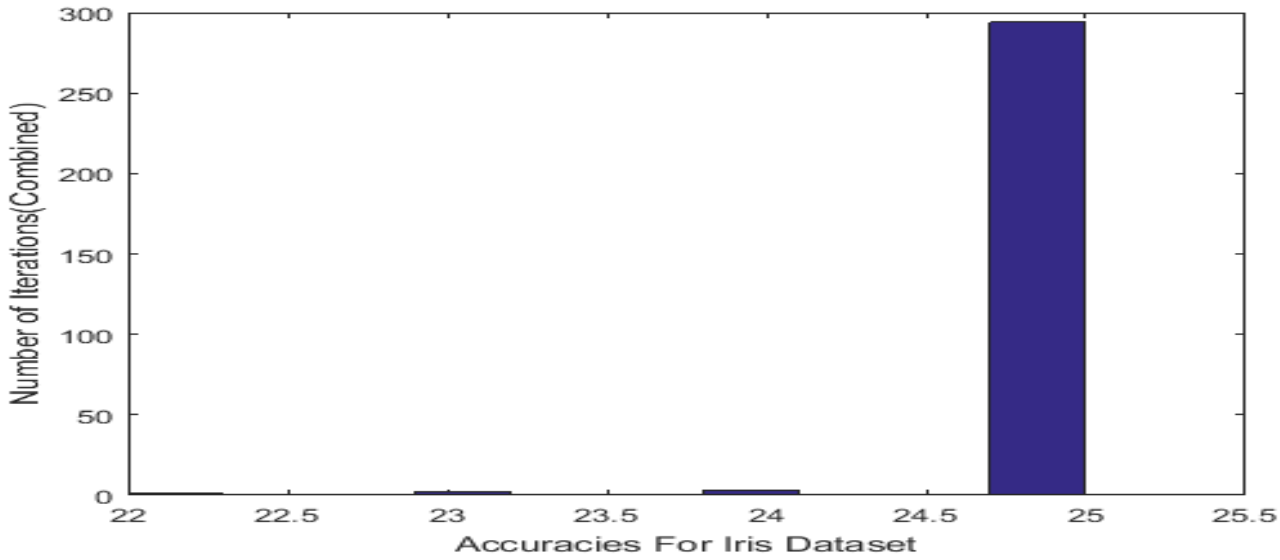

Iterations
Accuracies For Setosa
Iterations
Accuracies For Versicolor
Iterations
Accuracies For Virginica

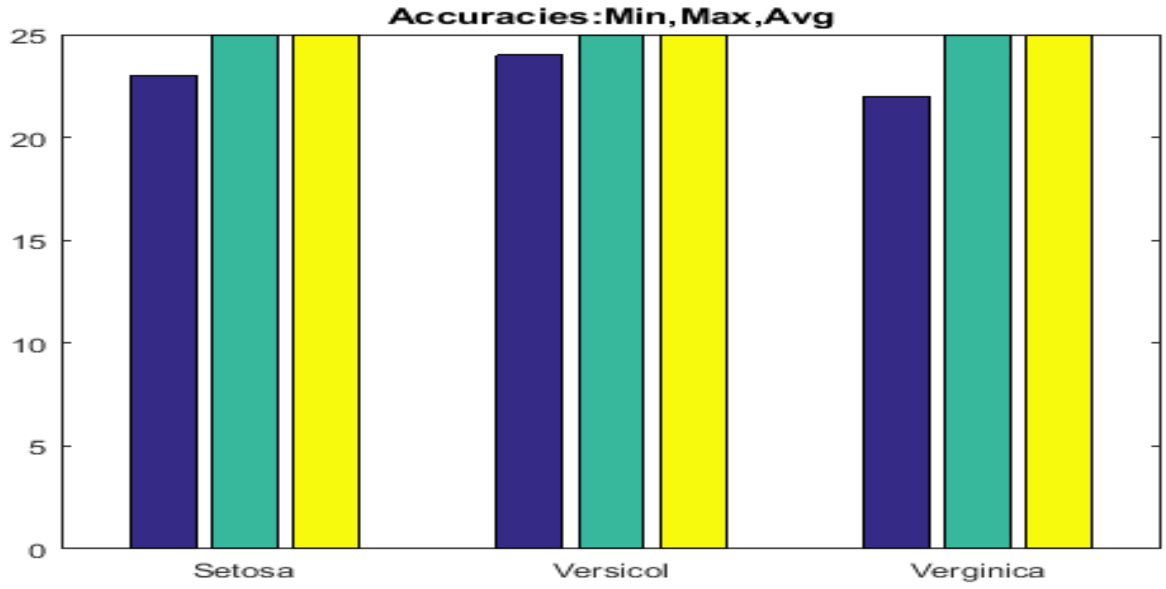

Number of Iterations(Combined)
Accuracies For Iris Dataset


Accuracies:Min,Max,Avg
Setosa
Versicol
Verginica

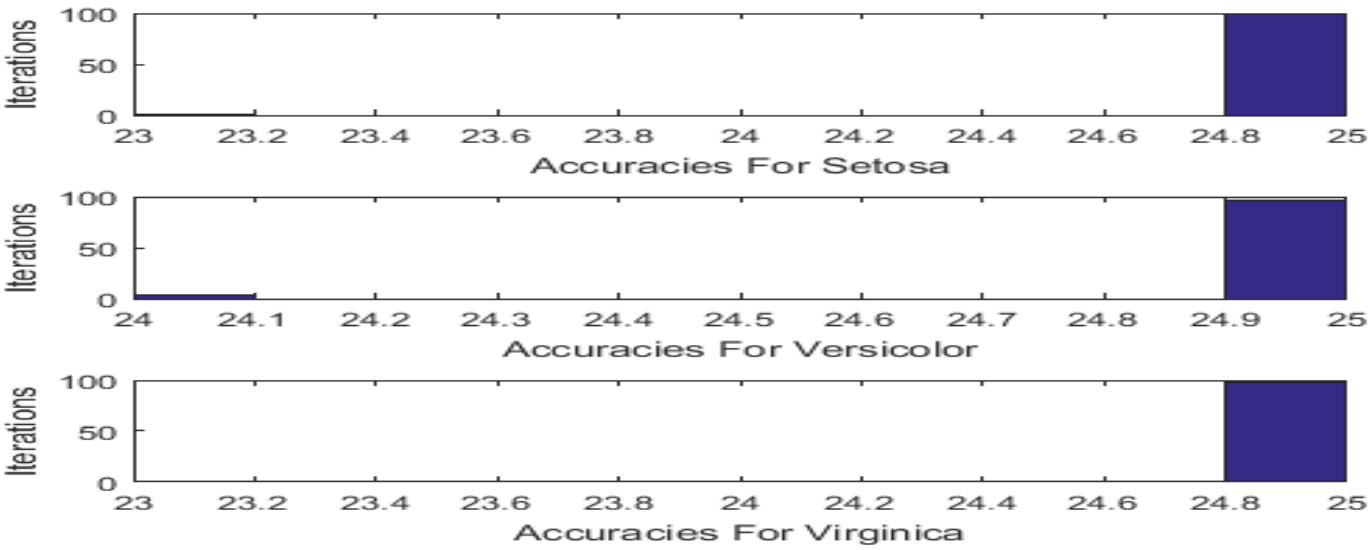
For K = 7:

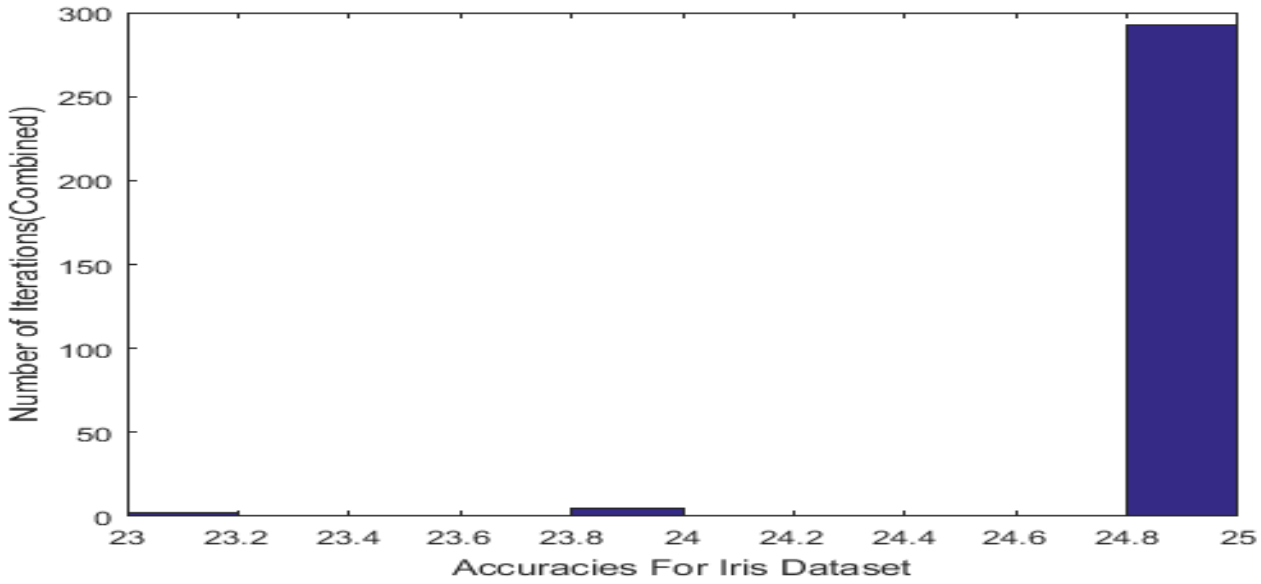

Iterations
100
50
0
23 23.2 23.4 23.6 23.8 24 24.2 24.4 24.6 24.8 25
Accuracies For Setosa
Iterations
100
50
0
24 24.1 24.2 24.3 24.4 24.5 24.6 24.7 24.8 24.9 25
Accuracies For Versicolor
Iterations
100
50
0
22 22.5 23 23.5 24 24.5 25 25.5
Accuracies For Virginica

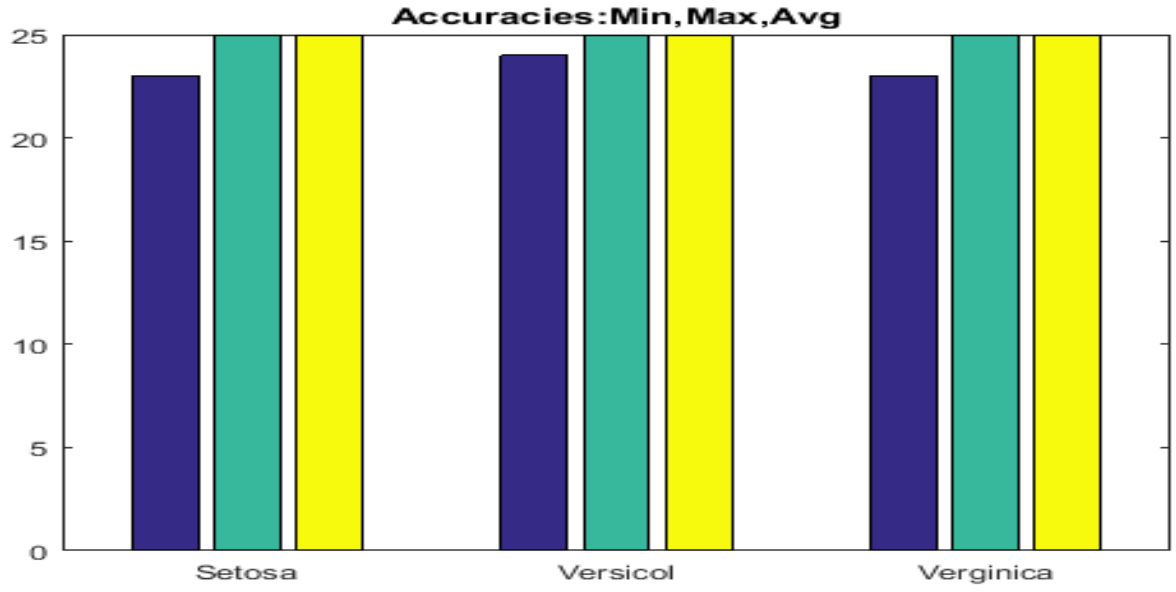

Number of Iterations(Combined)
300
250
200
150
100
50
0
22 22.5 23 23.5 24 24.5 25 25.5
Accuracies For Iris Dataset


Accuracies:Min,Max,Avg
25
20
15
10
5
0
Setosa
Versicol
Verginica

For K = 9:


Iterations
100
50
0
23
23.2
23.4
23.6
23.8
24
24.2
24.4
24.6
24.8
25
Accuracies For Setosa
Iterations
100
50
0
24
24.1
24.2
24.3
24.4
24.5
24.6
24.7
24.8
24.9
25
Accuracies For Versicolor
Iterations
100
50
0
23
23.2
23.4
23.6
23.8
24
24.2
24.4
24.6
24.8
25
Accuracies For Virginica


Number of Iterations(Combined)
300
250
200
150
100
50
0
23
23.2
23.4
23.6
23.8
24
24.2
24.4
24.6
24.8
25
Accuracies For Iris Dataset


Accuracies:Min,Max,Avg
25
20
15
10
5
0
Setosa
Versicol
Verginica