## **PAKISTAN COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH (PCSIR)**

## download

## **SUPPORT VECTOR MACHINE(SVM)**

## **NAME : AQSA ABDUL QADIR**

**POSITION: INTERNEE**

**DATE: 17-SEPT-2019**

## **Concepts Mapped:**

***1. What is SVM?***

***2.The ideology behind SVM.***

***3.Intuition development.***

***4.Terminologies used in SVM.***

***5. Hyperplane(Decision surface ).***

***6.Hard margin SVM.***

***7.Soft margin SVM.***

***8.Loss Function Interpretation of SVM.***

***9.Dual form of SVM.***

***10. What is Kernel trick?***

***11.Types of kernels.***

***12. Pros and cons of SVM.***

***13. Preparing data for SVM.***

***14. Conclusion***

# **Introduction:**

Support Vector Machine are perhaps one of the most popular and talked about machine learning algorithms.They were extremely popular around the time they were developed in the 1990s and continue to be the go-to method for a high performing algorithm with little tuning. In this blog we will be mapping the various concepts of SVC.

# **Abstract:**

Machine Learning is the wide technique , no dought , SVM is one of the machine learning technique which is use for learn the small data set . In this report we try to learn the iris data set by using SVM and we see that at 70% of training and 30% of test data set splitting we got the maximum learning upto 98% which is highly acceptable.

# **1. What is SVM?**

Support vector machines so called as SVM is a ***supervised learning algorithm*** which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). It is used for smaller dataset as it takes too long to process. In this set, we will be focusing on SVC.

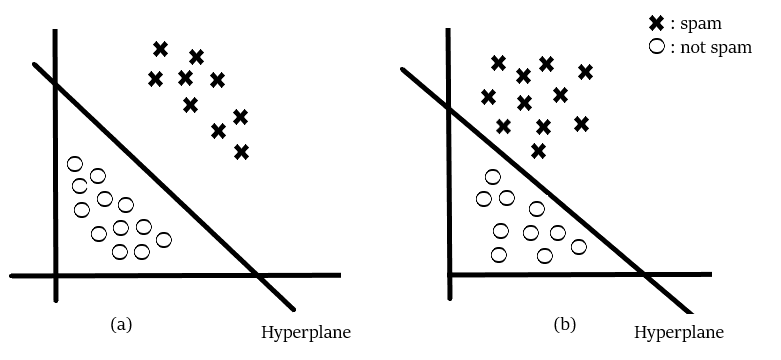
# **2. The ideology behind SVM:**

SVM is based on the idea of finding a hyperplane that best separates the features into different domains.

# **3. Intuition development:**

Consider a situation following situation:

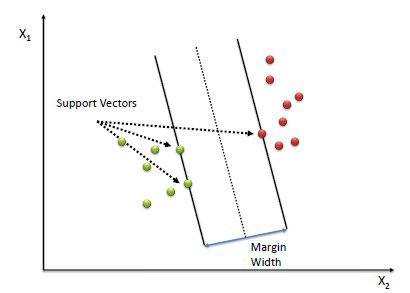
There is a stalker who is sending you emails and now you want to design a function( hyperplane ) which will clearly differentiate the two cases, such that whenever you received an email from the stalker it will be classified as a spam. The following are the figure of two cases in which the hyperplane are drawn, which one will you pick and why? take a moment to analyze the situation ……



I guess you would have picked the fig(a). Did you think why have you picked the fig(a)? Because the emails in fig(a) are clearly classified and you are more confident about that as compared to fig(b). Basically, SVM is composed of the idea of coming up with an ***Optimal hyperplane*** which will clearly classify the different classes(in this case they are binary classes).

# **4. Terminologies used in SVM:**

The points closest to the hyper plane are called as the ***support vector points*** and the distance of the vectors from the hyper plane are called the ***margins***.



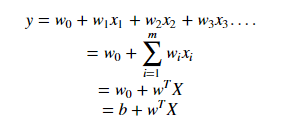
The basic intuition to develop over here is that more the farther SV points, from the hyper plane, more is the probability of correctly classifying the points in their respective region or classes. SV points are very critical in determining the hyper plane because if the position of the vectors changes the hyper plane’s position is altered. Technically this hyper plane can also be called as***margin maximizing hyper plane***.

# **5. Hyper plane(Decision surface ):**

For so long in this post we have been discussing the hyper plane, let’s justify its meaning before moving forward. The hyper plane is a function which is used to differentiate between features. In 2-D, the function used to classify between features is a line whereas, the function used to classify the features in a 3-D is called as a plane similarly the function which classifies the point in higher dimension is called as a hyper plane. Now since you know about the hyper plane lets move back to SVM.

Let’s say there are “m” dimensions:

thus the equation of the hyper plane in the ‘M’ dimension can be given as =



where,

Wi = vectors(W0,W1,W2,W3……Wm)

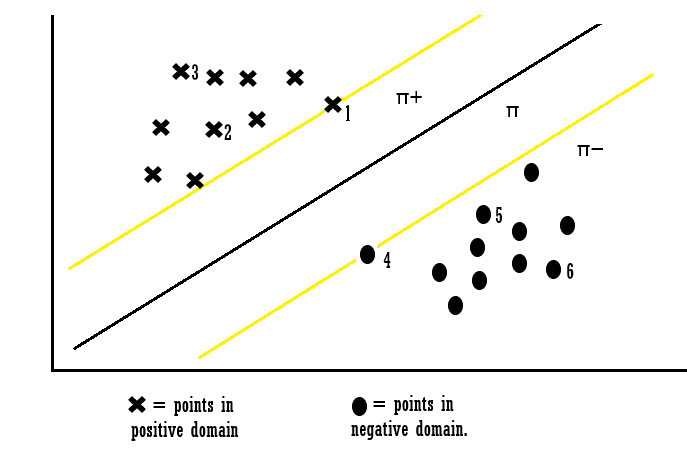
b = biased term (W0) (CHANGE ABLE TERM)

X = variables.

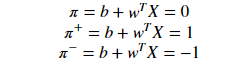
# **6. Hard margin SVM:**

Now,

*Assume 3 hyper planes namely (π, π+, π−) such that ‘π+’ is parallel to ‘π’ passing through the support vectors on the positive side and ‘π−’ is parallel to ‘π’ passing through the support vectors on the negative side.*



the equations of each hyper plane can be considered as:



for the point X1 :

IMG_267

*Explanation: when the point X1 we can say that point lies on the hyper plane and the equation determines that the product of our actual output and the hyper plane equation is 1 which means the point is correctly classified in the positive domain.*

for the point X3:

IMG_269

*Explanation: when the point X3 we can say that point lies away from the hyper plane and the equation determines that the product of our actual output and the hyper plane equation is greater 1 which means the point is correctly classified in the positive domain.*

for the point X4:

IMG_271

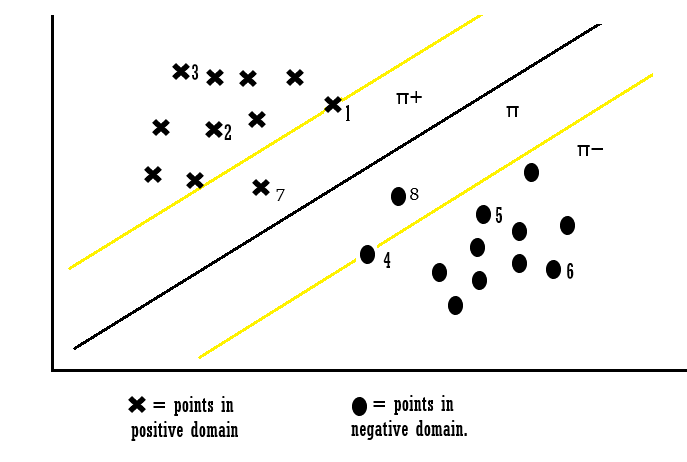
*Explanation: when the point X4 we can say that point lies on the hyper plane in the negative region and the equation determines that the product of our actual output and the hyper plane equation is equal to 1 which means the point is correctly classified in the negative domain.*

for the point X6 :

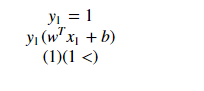
IMG_273

*Explanation: when the point X6 we can say that point lies away from the hyper plane in the negative region and the equation determines that the product of our actual output and the hyper plane equation is greater 1 which means the point is correctly classified in the negative domain.*

Let’s look into the constraints which are not classified:



for point X7:



*Explanation: When Xi = 7 the point is classified incorrectly because for point 7 the wT + b will be smaller than one and this violates the constraints. So we found the misclassification because of constraint violation. Similarly, we can also say for points Xi = 8.*

*Thus from the above examples, we can conclude that for any point Xi,*

***if Yi(WT\*Xi +b) ≥ 1:***

***then Xi is correctly classified***

***else:***

***Xi is incorrectly classified.***

So we can see that if the points are linearly separable then only our hyper plane is able to distinguish between them and if any outlier is introduced then it is not able to separate them. So these type of SVM is called***as hard margin SVM****(since we have very strict constraints to correctly classify each and every data point).*

# **7. Soft margin SVM:**

We basically consider that the data is linearly separable and this might not be the case in real life scenario. We need an update so that our function may skip few outliers and be able to classify almost linearly separable points. For this reason, we introduce a new ***Slack variable (***ξ ) which is called *Xi.*

if we introduce ξ it into our previous equation we can rewrite it as

IMG_279

Introduction of Xi

***if ξi= 0,***

***the points can be considered as correctly classified.***

***else:***

***ξi> 0 , Incorrectly classified points.***

so if ξi> 0 it means that Xi(variables)lies in incorrect dimension, thus we can think of ξi as an error term associated with Xi(variable). The average error can be given as;

IMG_281

average error

thus our objective, mathematically can be described as;

IMG_283

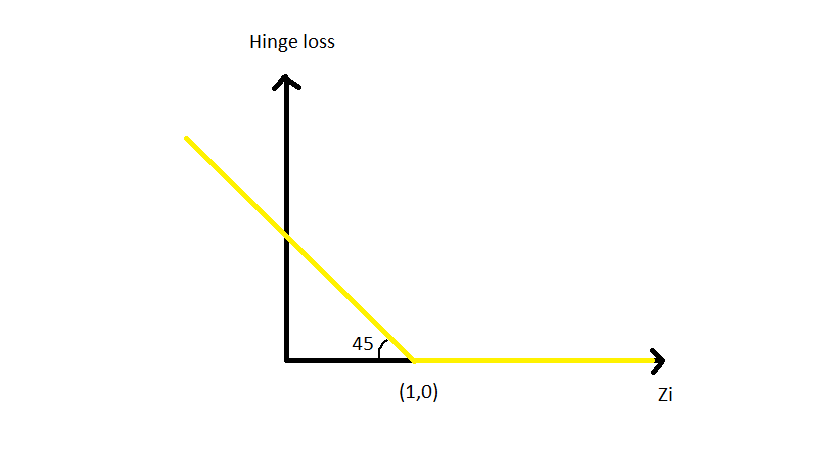
IMG_285

where ξi = *ςi*

***READING: To find the vector w and the scalar b such that the hyper plane represented by w and b maximizes the margin distance and minimizes the loss term subjected to the condition that all points are correctly classified.***

This formulation is called the Soft margin technique.

# **8. Loss Function Interpretation of SVM:**



IMG_289

when Zi is ≥ 1 then the loss is 0

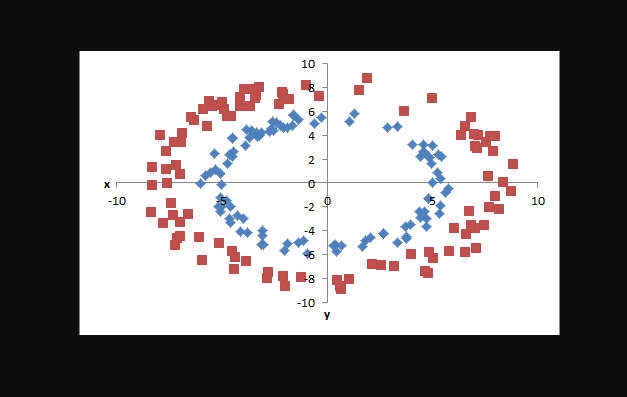
IMG_291

when Zi < 1 then loss increases.

thus it can be interpreted that hinge loss is max(0,1-Zi).

# **9. Dual form of SVM:**

Now, let’s consider the case when our data set is not at all linearly separable.



basically, we can separate each data point by projecting it into the higher dimension by adding relevant features to it as we do in logistic regression. But with SVM there is a powerful way to achieve this task of projecting the data into a higher dimension. The above-discussed formulation was the ***primal form of SVM*** . The alternative method is dual form of SVM which uses ***Lagrange’s multiplier*** to solve the constraints optimization problem.

IMG_295

IMG_297

***Note:***

***If αi>0 then Xi is a Support vector and when αi=0 then Xi is not a support vector.***

Observation:

1. To solve the actual problem we do not require the actual data point instead only the dot product between every pair of a vector may suffice.
2. To calculate the “b” biased constant we only require dot product.
3. The major advantage of dual form of SVM over Lagrange formulation is that it only depends on the ***α***.

# **10. What is Kernel trick?**

Coming to the major part of the SVM for which it is most famous, the***kernel trick***. The kernel is a way of computing the dot product of two vectors ****x**** and ****y**** in some (very high dimensional) feature space, which is why kernel functions are sometimes called “generalized dot product.

IMG_299

try reading this equation…

IMG_301

s.t = subjected to

Applying kernel trick means just to the replace dot product of two vectors by the kernel function.

# **11. Types of kernels:**

1. linear kernel
2. polynomial kernel
3. Radial basis function kernel (RBF)/ Gaussian Kernel

We will be focusing on the polynomial and Gaussian kernel since its most commonly used.

## **Polynomial kernel:**

In general, the polynomial kernel is defined as ;

IMG_303

b = degree of kernel & a = constant term.

in the polynomial kernel, we simply calculate the dot product by increasing the power of the kernel.

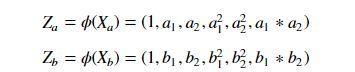
Example:

Let’s say originally X space is 2-dimensional such that

Xa = (a1 ,a2)

Xb = (b1 ,b2)

now if we want to map our data into higher dimension let’s say in Z space which is six-dimensional it may seem like



In order to solve the solve this dual SVM we would require the dot product of (transpose) Za ^t and Zb.

Method 1:

traditionally we would solve this by :

IMG_307

which will a lot of time as we would have to performs dot product on each datapoint and then to compute the dot product we may need to do multiplications Imagine doing this for thousand datapoints….

Or else we could simply use

Method 2:

using kernel trick:

IMG_309

In this method, we can simply calculate the dot product by increasing the value of power. Simple isn’t it?

## **Radial basis function kernel (RBF)/ Gaussian Kernel:**

Gaussian RBF(Radial Basis Function) is another popular Kernel method used in SVM models for more. RBF kernel is a function whose value depends on the distance from the origin or from some point. Gaussian Kernel is of the following format;

IMG_311

||X1 — X2 || = Euclidean distance between X1 & X2

Using the distance in the original space we calculate the dot product (similarity) of X1 & X2.

*Note: similarity is the angular distance between two points.*

## **Parameters:**

1. C: Inverse of the strength of regularization.

*Behavior: As the value of ‘c’ increases the model gets overfits.*

*As the value of ‘c’ decreases the model underfits.*

2. γ : Gamma (used only for RBF kernel)

*Behavior: As the value of ‘* γ*’ increases the model gets overfits.*

*As the value of ‘* γ*’ decreases the model underfits.*

# **12. Pros and cons of SVM:**

## **Pros:**

1. It is really effective in the higher dimension.
2. Effective when the number of features are more than training examples.
3. Best algorithm when classes are separable
4. The hyperplane is affected by only the support vectors thus outliers have less impact.
5. SVM is suited for extreme case binary classification.

## **cons:**

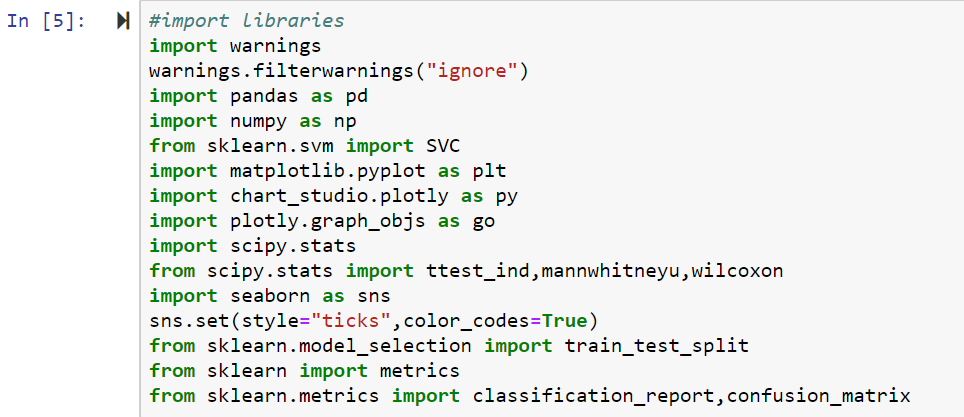
1. For larger dataset, it requires a large amount of time to process.
2. Does not perform well in case of overlapped classes.
3. Selecting, appropriately hyperparameters of the SVM that will allow for sufficient generalization performance.
4. Selecting the appropriate kernel function can be tricky.

## **PREPARING DATA FOR SVM:**

We use the iris data set to understand the SVM . We apply SVM on the starting 5 rows and columns.

**IMPORT LIBRARY:**

#import libraries  
import warnings  
warnings.filterwarnings(“ignore”)  
import pandas as pd  
import numpy as np  
from sklearn.svm import SVC  
import matplotlib.pyplot as plt  
import plotly.plotly as py  
import plotly.graph\_objs as go  
import scipy.stats  
from scipy.stats import ttest\_ind,mannwhitneyu,wilcoxon  
import seaborn as sns  
sns.set(style=”ticks”,color\_codes=True)  
from sklearn.model\_selection import train\_test\_split  
from sklearn import metrics  
from sklearn.metrics import classification\_report,confusion\_matrix



**LOAD DATA:**

#load data set URL  
url = “https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data”  
names = [‘sepal-length’, ‘sepal-width’, ‘petal-length’, ‘petal-width’, ‘class’]  
data = pd.read\_csv(url, names=names)

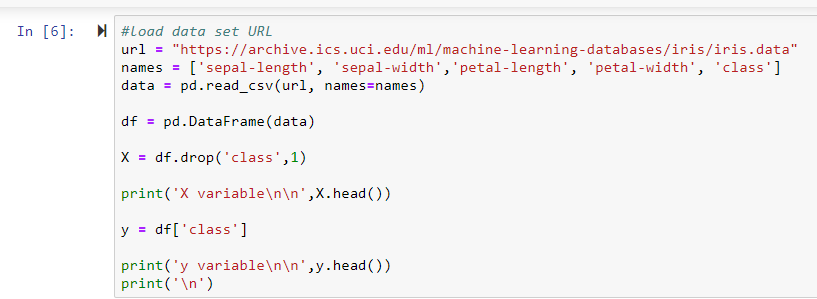
df = pd.DataFrame(data)

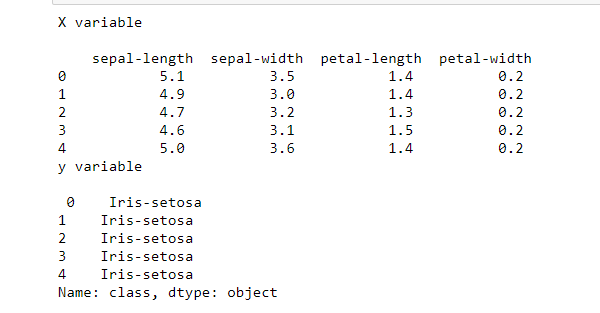
X = df.drop(‘class’,1)

print(“X variable\n\n”,X.head())

y = df[‘class’]

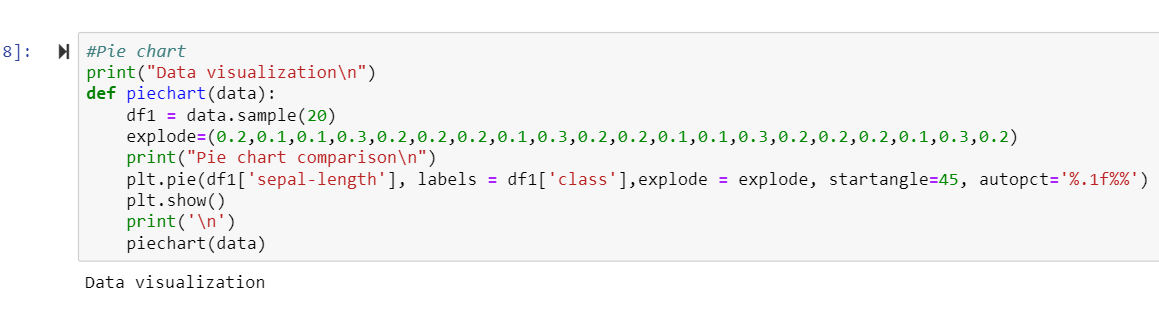
print(“y variable\n\n”,y.head())  
print(“\n”)





**PIE CHART FUNCTION:**

#Pie chart  
print(“Data visualization\n”)  
def piechart(data):  
df1 = data.sample(20)  
explode=(0.2,0.1,0.1,0.3,0.2,0.2,0.2,0.1,0.3,0.2,0.2,0.1,0.1,0.3,0.2,0.2,0.2,0.1,0.3,0.2)  
print(“Pie chart comparison\n”)  
plt.pie(df1[‘sepal-length’], labels = df1[‘class’],explode = explode, startangle=45, autopct=’%.1f%%’)  
plt.show()  
print(“\n”)  
piechart(data)



**GRAPH PLOTTING:**

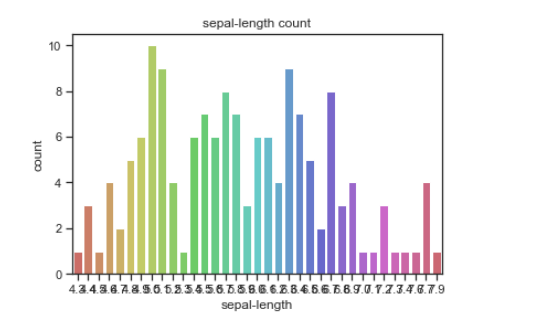
#Count plot  
print(“Count plots”)  
sns.countplot(x=df[‘sepal-length’],data=df, palette=’hls’)  
plt.title(“sepal-length count”)  
plt.show()  
print(“\n”)

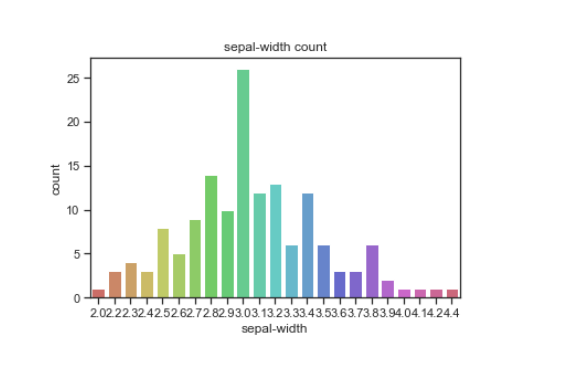
sns.countplot(x=df[‘sepal-width’],data=df, palette=’hls’)  
plt.title(“sepal-width count”)  
plt.show()  
print(“\n”)

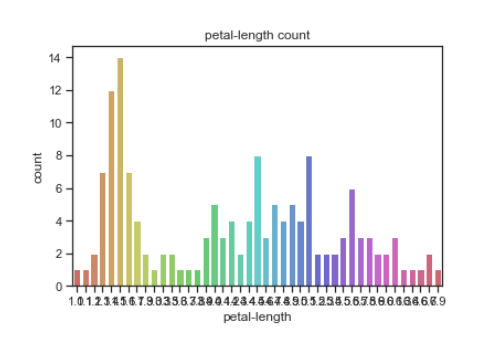
sns.countplot(x=df[‘petal-length’],data=df, palette=’hls’)  
plt.title(“petal-length count”)  
plt.show()  
print(“\n”)

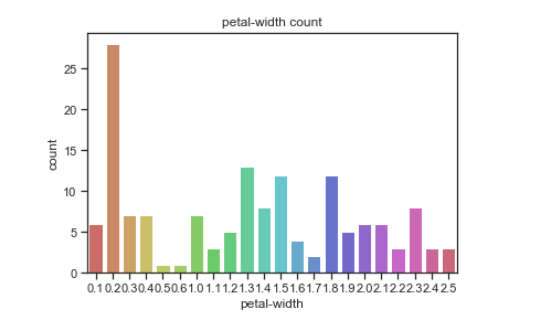
sns.countplot(x=df[‘petal-width’],data=df, palette=’hls’)  
plt.title(“petal-width count”)  
plt.show()  
print(“\n”)







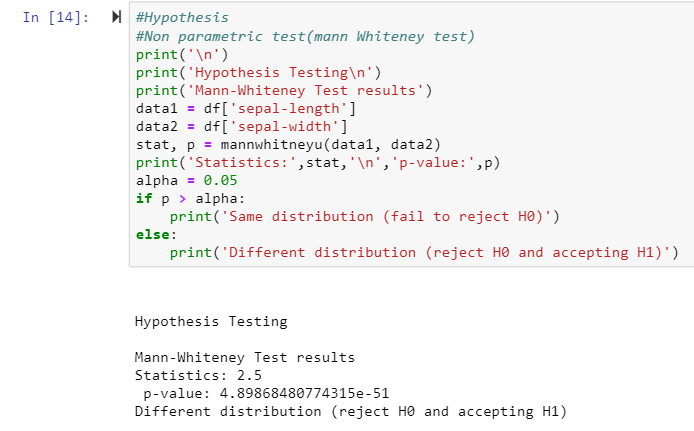


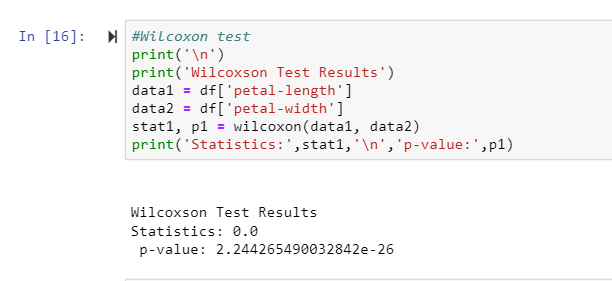


**STATISTIC CHECK:**

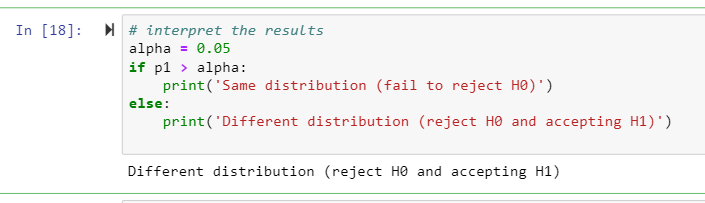
#Hypothesis  
#Non parametric test(mann Whiteney test)  
print(“\n”)  
print(“Hypothesis Testing\n”)  
print(“Mann-Whiteney Test results”)  
data1 = df[‘sepal-length’]  
data2 = df[‘sepal-width’]  
stat, p = mannwhitneyu(data1, data2)  
print(“Statistics:”,stat,”\n”,”p-value:”,p)  
alpha = 0.05  
if p > alpha:  
print(‘Same distribution (fail to reject H0)’)  
else:  
print(‘Different distribution (reject H0 and accepting H1)’)

#Wilcoxon test  
print(“\n”)

  
print(“Wilcoxson Test Results”)  
data1 = df[‘petal-length’]  
data2 = df[‘petal-width’]  
stat1, p1 = wilcoxon(data1, data2)  
print(“Statistics:”,stat1,”\n”,”p-value:”,p1)



# interpret the results  
alpha = 0.05  
if p1 > alpha:  
print(‘Same distribution (fail to reject H0)’)  
else:  
print(‘Different distribution (reject H0 and accepting H1)’)  
print(“\n”)



**SPLITT DATA INTO TEST AND TRAINING:**

#Split the data train and test  
print(“Support Vector Machine Results\n”)  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=0)

**SUPPORT VECTOR MACHINE:**  
#Support Vector machine classifier

svclassifier = SVC(kernel=’linear’)

#fit the model  
svclassifier.fit(X\_train, y\_train)

#predict the test data  
y\_pred = svclassifier.predict(X\_test)

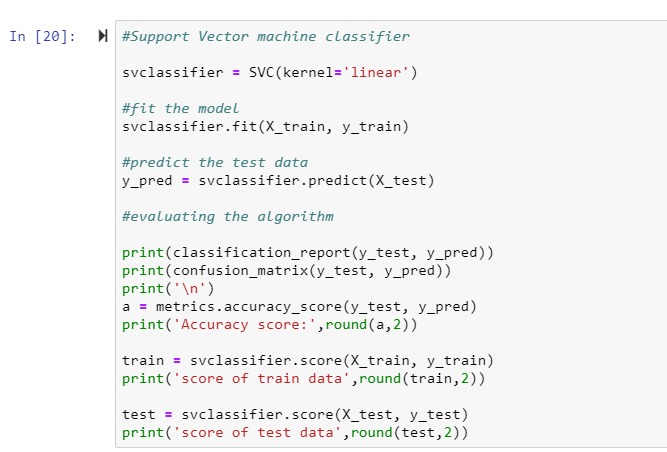
**EVALUATION:**

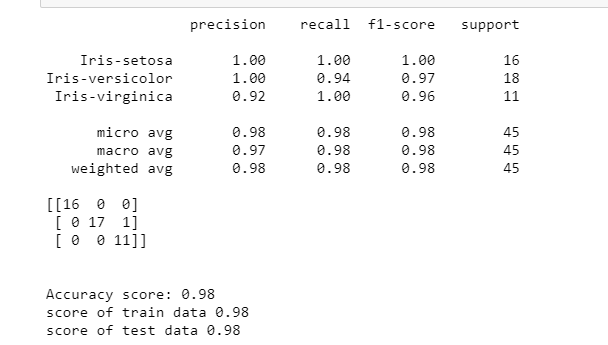
#evaluating the algorithm

print(classification\_report(y\_test, y\_pred))  
print(confusion\_matrix(y\_test, y\_pred))  
print(“\n”)  
a = metrics.accuracy\_score(y\_test, y\_pred)  
print(“Accuracy score:”,round(a,2))

train = svclassifier.score(X\_train, y\_train)  
print(“score of train data”,round(train,2))

test = svclassifier.score(X\_test, y\_test)  
print(“score of test data”,round(test,2))





1. **CONCLUSION:**

By the all upper studies we can say that support vector machine(SVM) technique used for smallest data sets . It can give best result on the smallest data set rather than big data sets.