Himanshu Kumar

# WRITE A PYTHON PROGRAM USING SCIKIT-LEARN TO SPLIT THE IRIS DATASET INTO 80% TRAIN DATA AND 20% TEST DATA. OUT OF TOTAL 150 RECORDS, THE TRAINING SET WILL CONTAIN 120 RECORDS AND THE TEST SET CONTAINS 30 OF THOSE RECORDS. TRAIN OR FIT THE DATA INTO THE MODEL AND CALCULATE THE ACCURACY OF THE MODEL USING THE K NEAREST NEIGHBOUR ALGORITHM.

**Answer:**

# QUERY:

import numpy as np import pandas as pd import seaborn as sns

idf=sns.load\_dataset("Iris") X=idf.drop(['species'], axis='columns') y=idf.species

from sklearn.model\_selection import train\_test\_split #Adding Train Test Split Model

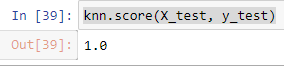
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=1)

len(X\_train) len(X\_test)

from sklearn.neighbors import KNeighborsClassifier #Importing KNN Classifier

knn = KNeighborsClassifier(n\_neighbors=8) knn.fit(X\_train, y\_train) knn.score(X\_test, y\_test)

# OUTPUT:



1. **FURTHER TRAIN OR FIT INTO THE MODEL AND CALCULATE THE PERFORMANCE FOR DIFFERENT VALUES OF K? (USE IRIS DATASET FROM SK LEARN DATASETS)**

# Answer:

**QUERY:**

import numpy as np import pandas as pd import seaborn as sns

idf=sns.load\_dataset("Iris") X=idf.drop(['species'], axis='columns') y=idf.species

from sklearn.model\_selection import train\_test\_split #Adding Train Test Split Model

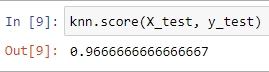
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=1)

len(X\_train) len(X\_test)

from sklearn.neighbors import KNeighborsClassifier #Importing KNN Classifier

knn = KNeighborsClassifier(n\_neighbors=12) knn.fit(X\_train, y\_train) knn.score(X\_test, y\_test)

# OUTPUT:



knn = KNeighborsClassifier(n\_neighbors=5)**#K value=5**

knn.fit(X\_train, y\_train) knn.score(X\_test, y\_test)



knn = KNeighborsClassifier(n\_neighbors=20) **#K value=20**

knn.fit(X\_train, y\_train) knn.score(X\_test, y\_test)



# FURTHER PERFORM LOGISTICS REGRESSION CLASSIFICATION ON THE ABOVE DATASET AND COMPARE THE ACCURACY BETWEEN TWO ALGORITHMS?

**Answer:**

# QUERY:

from sklearn.linear\_model import LogisticRegression Log=LogisticRegression()

Log.fit(X\_train, y\_train) Log.score(X\_test, y\_test)

# OUTPUT:

1. **WHAT IS THE DIFFERENCE BETWEEN PARAMETRIC & NON PARAMETRIC LEARNING? GIVE EXAMPLES**

# Answer:

**Parametric Test Definition**

In Statistics, a parametric test is a kind of the hypothesis test which gives generalizations for generating records regarding the mean of the primary/original population. The t-test is carried out based on the students t-statistic, which is often used in that value.

The t-statistic test holds on the underlying hypothesis which includes the normal distribution of a variable. In this case, the mean is known, or it is considered to be known. For finding the sample from the population, population variance is identified. It is hypothesized that the variables of concern in the population are estimated on an interval scale.

**Example:** Logistic Regression, Naïve Bayes Model, etc.

# Non-Parametric Test Definition

The non-parametric test does not require any population distribution, which is meant by distinct parameters. It is also a kind of hypothesis test, which is not based on the underlying hypothesis. In the case of the non-parametric test, the test is based on the differences in the median. So, this kind of test is also called a distribution-free test. The test variables are determined on the nominal or ordinal level. If the independent variables are non-metric, the non-parametric test is usually performed.

**Example:** KNN, Decision Tree Model, etc.

# WHAT IS THE DIFFERENCE BETWEEN SUPERVISED LEARNING AND UN SUPERVISED LEARNING? GIVE EXAMPLES

**Answer:**

**Supervised Learning:** [Supervised learning](https://www.ibm.com/cloud/learn/supervised-learning) is a machine learning approach that’s defined by its use of labeled datasets. These datasets are designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time. Supervised learning can be separated into two types of problems when [data mining](https://www.ibm.com/cloud/learn/data-mining): classification and regression.

**Unsupervised Learning:** [Unsupervised learning](https://www.ibm.com/cloud/learn/unsupervised-learning) uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are “unsupervised”). Unsupervised learning models are used for three main tasks: clustering, association and dimensionality reduction.

# EXPLAIN VARIOUS STEPS INVOLVED IN MACHINE LEARNING PROCESS?

**Answer:** The machine learning process involves the following steps:

1. **Gathering Data:** In this step, a certain type of data is collected from the sources. This is the foundation of machine learning process.
2. **Preparing the Data:** Once the data is gathered, the next step is to prepare it for further processes. Skewness is also checked for biasness of the data.
3. **Choosing a Model:** A well suited data model is chosen with different goals in mind. For example some models are good fit for text while some are good with images.
4. **Training a model:** This is the heart of a machine learning model, lots of learning is done at this stage.
5. **Evaluation:** After the model is trained, it need to be tested, how it work in a real world scenarios. It is used to check the model’s proficiency.
6. **Hyperparemeter Tuning:** When the evaluation is successful, this step tries to improve the positive results.
7. **Prediction:** This is the stage where we consider the model to be ready for implementation.

# PERFORM DIMENSIONALITY REDUCTION USING PCA ON THE US ARRESTS DATASET (ENCLOSED HEREWITH). WHAT VARIANCE CAN BE EXPLAINED BY PC1 & PC2

**Answer:**

# QUERY:

import pandas as pd

from sklearn.preprocessing import StandardScaler from sklearn.decomposition import PCA

import seaborn as sns arrestdf=pd.read\_csv("[Template] USArrests .csv.csv")

arrestdf.rename(columns={'Unnamed: 0':'Country'}, inplace=True) arrestdf.drop("Country", axis=1, inplace=True)

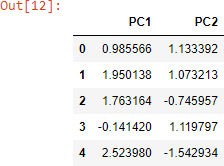
arrestdf.head() scaler=StandardScaler()

s\_data=pd.DataFrame(scaler.fit\_transform(arrestdf)) s\_data

pca=PCA(n\_components=2) pca.fit(s\_data) data\_pca=pca.transform(s\_data)

data\_pca=pd.DataFrame(data\_pca, columns=['PC1', 'PC2']) data\_pca.head()

# OUTPUT:



data\_pca.var().round(2)

# OUTPUT:



**Q8 Create Basic association rule manually. The 'database' below has four transactions. What association rules can be found in this set, if the minimum support (i.e coverage) is 60% and the minimum confidence (i.e., accuracy) is 80%**

# Trans\_id Item list T1 {K, A, D, B}

**T2 {D, A C, E, B}**

# T3 {C, A, B, E}

**T4 {B, A, D}**

# Q9 Create Association Rules using Apriori algorithm using the dataset ‘apriori data’) attached herewith.

**Q10 Show how the count of rules vary by changing Support & Confidence thresholds.**

# Q11 WHAT IS MEANT BY COLLABORATIVE FILTERING IN RECOMMENDER SYSTEM?

**Answer:** Collaborative filtering is the predictive process behind recommendation engines. Recommendation engines analyze information about users with similar tastes to assess the probability that a target individual will enjoy something, such as a video, a book or a product. Collaborative filtering is also known as social filtering. Collaborative filtering uses algorithms to filter data from user reviews to make personalized recommendations for users with similar preferences. Collaborative filtering is also used to select content and advertising for individuals on social media. Three types of collaborative filtering commonly used in recommendation systems are neighbor-based, item-to-item and classification- based.

# Q12 WHICH OF THE FOLLOWING IS A REASONABLE WAY TO SELECT THE NUMBER OF PRINCIPAL COMPONENTS "K"?

1. **Choose k to be the smallest value so that at least 99% of the variance is retained.**

# Choose k to be 99% of m (k = 0.99\*m, rounded to the nearest integer).

1. **Choose k to be the largest value so that 99% of the variance is retained.**

# Use the elbow method.

**Answer:** B. Choose k to be 99% of m (k = 0.99\*m, rounded to the nearest integer).

# Q13 WHICH OF THE FOLLOWING STATEMENT IS FALSE IN THE CASE OF THE KNN ALGORITHM?

1. **For an exceptionally large value of K, points from other classes may be included in the neighbourhood.**

# For the exceedingly small value of K, the algorithm is overly sensitive to noise.

1. **KNN is used only for classification problem statements.**

# KNN is a lazy learner.

**Answer:** (C) KNN is used only for classification problem statements.

# Q14 WHICH OF THE FOLLOWING STATEMENT IS TRUE.

1. **Outliers should be identified and removed always from the dataset.**

# Outliers can never be present in the testing dataset.

1. **Outlier is a data point which is significantly close to other data points.**

# The nature of our business problem determines how outliers are used.

**Answer:** D) The nature of our business problem determines how outliers are used.

# Q15 IN REGRESSION MODELLING WE DEVELOP A MATHEMATICAL EQUATION THAT DESCRIBES HOW, (PREDICTOR-INDEPENDENT VARIABLE, RESPONSE-DEPENDENT VARIABLE)

1. **one predictor and one or more response variables are related.**

# several predictors and several response variables response is related.

1. **one response and one or more predictors are related.**

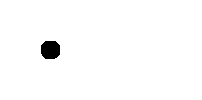
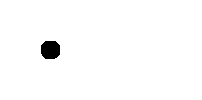
# All of these are correct.

**Answer:** (A) one predictor and one or more response variables are related.

# Q16 WHEN PERFORMING REGRESSION OR CLASSIFICATION, WHICH OF THE FOLLOWING IS THE CORRECT WAY TO PRE-PROCESS THE DATA?

1. **Normalize the data PCA training.**

# PCA Normalize PCA output training.



1. **Normalize the data PCA normalize PCA output training.**

# None of the above.

**Answer:** A) Normalize the data PCA training.