**Assignment -2 CS-805 Course Machine Learning and Data Mining Name: Robert Akinie**

1. How different is this book against many other slower paced books available outside?
2. It does not encourage you how to become good at Python programming and syntax before working on a ML project
3. It does not delve into the underlying theory and parameters for each ML algorithm before usage
4. It walks users through the process of building predictive ML models in Python
5. In a Machine Learning (ML) Project, What are the 9 processes/tasks used for predictive modelling ?
6. Load Dataset
7. Preprocess the dataset, like data cleaning
8. Determine Descriptive Statistics of the dataset
9. Data Visualization
10. Feature Engineering
11. Create training and test datasets
12. Build model based on training dataset
13. Evaluate model using methods like Confusion Matrix, precision, recall and f1 score
14. Validation of dataset
15. Prediction of model using test dataset
16. The data sets used here are obtained from which repository, and what are the three popular datasets that are used for three different tasks of machine learning. Also write down the categories of ML problems ?

**UCI Machine Learning repository.**

**Iris flowers dataset, Boston House Price dataset, and Sonar dataset.**

**Three categories of ML are classification, regression and clustering.**

1. How are Recipes different than a Tutorial WorkBook ?

**Since recipes are code snippets rather than tutorials, they provide just enough code to work, and they are demonstrative, not exhaustive. They are small examples that show users howe to do a specific example**

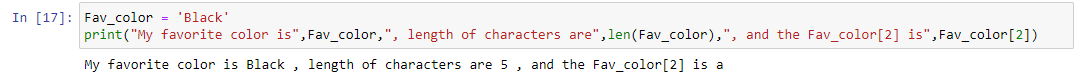
1. Describe the SciPy Ecosystem and its 3 popular Libraries? What are these Libraries used for ?

**The Scipy ecosystem is an ecosystem of librariese for science, mathematics and engineering. Popular libraries are numpy, pandas and matplotlib. Numpy is best for working with arrays, pandas is a library for the organization and analysis of data using tools and data structures, and matplotlib is a library that allows for the creation of two dimensional charts and plots from data**

1. Describe the SciKit-learn Ecosystem and its relationship with SciPy ToolKit?

**Scikit-learn library is how users ML experts develop and practice ML in Python. It is a ScPy plug-in and is built on the latter.**

1. Print the Variable ‘Fav\_Color’ with your favorite Color name, Size of the characters in the Color, and print the character in “Fav\_Color[2]” ? Refer the Eg. Listing # 3.1



1. Print the Variables “Height in Foot-and-Inches” and your “Lucky digit” ? What would be the data types and how is it different with initialization in Python against C-Language ? Hint: Listing # 3.3

Text, letter

Description automatically generated

1. Print Boolean Variable Hair\_Black = True ?

Graphical user interface, text, application, chat or text message

Description automatically generated

1. Assign multiple array elements such as X,Y,Z as 10, 11, and 12 ?

Graphical user interface, text, application

Description automatically generated

1. The Variable “Room\_Num” has value of None. Write an if else code to print your room Number if provided or None else not provided ? Hint:-- Listing # 3.11  
    Graphical user interface, text, application

   Description automatically generated
2. Highlight what is wrong in the below code ?

speed = 60

if speed == 100:

print 'That is fast'

elseif speed > 200;

print 'That is too fast'

else:

print 'That is safe'

**There is no indentation block for the if-else statements.**

**The else-if statement syntax is invalid: elif :**

**The print statements have no parentheses for their arguments**

1. Highlight what is wrong in the below code ?

speed = 250

if (speed == 100):

print 'That is fast'

elif speed > 200;

print 'That is too fast'

else speed <250

print 'That is safe'

**There is no indentation block for the if-else statements.**

**The else-if statement syntax is invalid: elif :**

**The else statement has invalid syntax: else speed <250:**

**The print statements have no parentheses for their arguments**

1. Using a For Loop, print the variable X from 10 to 16 by incrementing 1 in the range ? Listing # 3.13.

Graphical user interface, text

Description automatically generated

1. Using a While Loop, print the variable Z from 10 to 16 by incrementing 1 in the range ? Listing # 3.15.

Text

Description automatically generated with medium confidence

1. Write the Code for creating a list/array of 5 values between 5 and 10, and add one more 6th value between 20 and 25. Print the 2nd Value of the list, length of the list, and the total list in the output ?

Graphical user interface, text, application

Description automatically generated

1. Write the code for following tasks: 1) Create a Dictionary of Key-Value Pairs for “X, Y and Z” with variables any between 20 to 25. (2) Print the Value of KEY- Y, (3) Change the Value of KEY-X and Print it, (4) Print the Keys List, (5) Print the Values List, (6) Use the For Loop to print the Complete List ? Hint # Listing # 3.21

Text, letter

Description automatically generated

1. Repeat Question # 17 with Creating an Array Variable ‘Choice\_List’ instead of Dictionary and print all the above items asked in Question # 17 ? Hint: Listing # 3.25

Text

Description automatically generated

1. Make a line plot of the Choice\_List created in Question # 18 and paste the graph here ?

Chart, line chart

Description automatically generated

1. Repeat Question # 18 and provide a Scatter Plot here ?

Chart

Description automatically generated

Part-II :

1. Create a Function Named Area\_House ? with variables Length =150ft and Breadth = 100ft. The function Area\_House should accept two arguments for length and breadth, and shall print the result from inside the function. ? Hint:-- See Listing # 3.23

Graphical user interface, text, application

Description automatically generated

1. Create an array variable my\_list of size [3 rows ,4 columns] with numbers from 1 to 20? And then print The 2nd row; last row; value of my\_list[2,3]; and 4th full column ? Hint: Listing # 3.28

Graphical user interface, text, application

Description automatically generated

1. Using my\_list created in Question # 2,
2. add all the values in the row 1 and print the result
3. multiply each value in col. 2 against col 3. and print the result in a column again ?

Text

Description automatically generated

1. Create and Print the DataFrame variable my\_matrix with rows names as (x,y,z,) and Column names as (a,b,c,d) and fill it with the values from the my\_list ? Hint : Listing # 3.37

Graphical user interface, application

Description automatically generated

1. Download the Data Iris.CSV from the UCI Machine Learning Repository. Use the appropriate Libraries and Load the data in the variable My\_Dataset with Column names and
2. print the sample data of first 5 rows along with the column names.

A picture containing text

Description automatically generated

1. Print the Size of the above My\_Dataset ? Listing # 4.3 ?

Graphical user interface

Description automatically generated with low confidence

1. Print the data types of all the four columns in the My\_Dataset ?

Graphical user interface, text

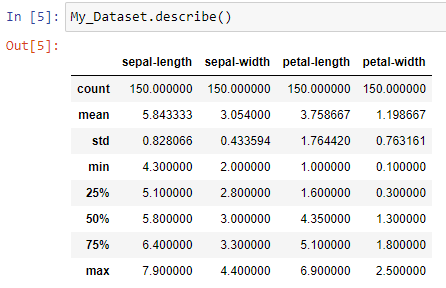
Description automatically generated

1. Print the distribution size of samples for four different classes of My\_Dataset?

Text

Description automatically generated with medium confidence

1. Print the Descriptive Statistics of the above My\_Dataset; such as mean, std. dev., Count, Min Value, Max Value, Median



1. What is skewness of a data distribution ? Compute the skewness of each column of My\_Dataset and print it ?

**Skew refers to a normal/Gaussian distribution that has been shifted in one direction.**

Graphical user interface, text

Description automatically generated

1. Create the probability density plots of all the four attributes of the My\_Dataset ?

Chart

Description automatically generated

1. Verify that the Skewness and Density plots are agreeing with each other? Explain by taking an example?

**Yes, they agree with each other. Taking a look at the sepal-length feature, the skew is positive, and from the density plot, the tail of the distribution is on its right side. Also, the median is lower than the mean.**

1. Compute the Correlation coefficients matrix of the attributes and also show the correlation matrix plot of My\_Dataset ?

Chart

Description automatically generated

1. Create a Scatter Plot Matrix for the My\_Dataset with Labels on X-axis and Y-axis and explain what you can infer in comparison of Scatter Plots and Correlation plots ?

**Chart, scatter chart

Description automatically generated**

**Both plots show degree of correlation between two features, but scatter matrix shows it in greater detail. Correlation matrix plots gives a basic idea about the strength of two features, essentially a linear relationship, whilst scatter matrix plot displays the strength, direction and form of that relationship.**

**Appendix**

import random

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from pandas import read\_csv

#Part 1

print("Question 7")

Fav\_color = 'Black'

print("My favorite color is", Fav\_color, "length of characters are", len(Fav\_color), "and the Fav\_color[2] is", Fav\_color[2])

print("\nQuestion 8")

Height\_ft = '6 ft'

Lucky\_digit = 7

print("Height is", Height\_ft, "and lucky digit is", Lucky\_digit)

print(type(Height\_ft), type(Lucky\_digit))

print("\nQuestion 9")

Hair\_black = True

print(Hair\_black)

print("\nQuestion 10")

X, Y, Z = 10, 11, 12

print(X, Y, Z)

print("\nQuestion 11")

Room\_Num = None

if Room\_Num:

print(Room\_Num)

else:

print("None provided")

print("\nQuestion 14")

for i in range(10, 17):

print(i)

print("\nQuestion 15")

i = 10

while i < 17:

print(i)

i+=1

print("\nQuestion 16")

list1 = [random.randint(6, 9) for i in range(0, 5)]

list1.append(random.randint(20, 25))

print("Second list value: %d" %list1[1], "\nList length: %d" %len(list1))

print(list1)

print("\nQuestion 17")

dict1 = {'X':21, 'Y':23, 'Z': 24}

print("dict1 is ", dict1, "\nY value is: %d" %dict1['Y'])

dict1['X'] = 100

print("X new value is: %d" %dict1['X'])

print(dict1.keys(), dict1.values())

for (k, v) in dict1.items():

print(k, v)

print("\nQuestion 18")

Choice\_List = ([['X', 'Y', 'Z'], [21, 23, 24]])

print("Choice\_List is", Choice\_List,"\nY value is: ",Choice\_List[1][1])

Choice\_List[1][0] = 100

print("X new value is: ", Choice\_List[1][0])

print("Keys are: ",Choice\_List[0], "and values are: ", Choice\_List[1])

for i in Choice\_List:

print(i)

plt.title("Question 19")

plt.plot(Choice\_List[1])

plt.xlabel('keys')

plt.ylabel('Choice\_List')

plt.show()

x = (0, 1, 2)

plt.scatter(x, Choice\_List[1])

plt.title("Question 20")

plt.show()

#Part 2

print("\nQuestion 21")

l=150

b=100

def Area\_House(length, breadth):

return length\* breadth

print(Area\_House(l, b))

print("\nQuestion 22")

my\_list = np.array([[1, 2, 3, 4], [5, 6, 7, 8], [9, 10, 11, 12]])

print(my\_list,"\n2nd row is: ", my\_list[1], "\nValue of my\_list[2, 3] is: ", my\_list[2,3], "\nLast column is: ", my\_list[:,3])

print("\nQuestion 23")

result = 0

for i in my\_list[0]:

result+=i

print(result)

column = my\_list[:, 1]\*my\_list[:, 2]

print(column)

print("\nQuestion 24")

col = ['a', 'b', 'c', 'd']

ro = ['x', 'y', 'z']

my\_matrix = pd.DataFrame(my\_list, index = ro, columns = col)

print(my\_matrix)

print("\nQuestion 25-30")

url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"

names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']

My\_Dataset = read\_csv(url, names=names)

print("\nSample Data is: ")

print(My\_Dataset.head())

print("\nSize of dataset is: ")

print(My\_Dataset.shape)

print("\nData types of my\_Dataset is: ")

print(My\_Dataset.dtypes)

print("\nDistribution size of samples: ")

print(My\_Dataset.groupby('class').size())

print("\nDescriptive statistics: ")

print(My\_Dataset.describe())

print("\nSkews of My\_Dataset: ")

print(My\_Dataset.skew())

My\_Dataset.plot(kind='density', subplots=True, layout=(2,2), sharex=False)

plt.title("Question 27")

plt.show()

print("\nCorrelations of My\_Dataset: ")

print(My\_Dataset.corr())

My\_Dataset.corr(method = 'pearson')

fig = plt.figure()

ax=fig.add\_subplot(111)

cax=ax.matshow(My\_Dataset.corr(), vmin=-1, vmax=1)

fig.colorbar(cax)

ticks = np.arange(0,5,1)

ax.set\_xticks(ticks)

ax.set\_yticks(ticks)

ax.set\_xticklabels(names)

ax.set\_yticklabels(names)

plt.title("Question 29")

plt.show()

pd.plotting.scatter\_matrix(My\_Dataset)

plt.title("Question 30")

plt.show()