# National University of Computer and Emerging Sciences

**FAST School of Computing        Spring-2023        Islamabad Campus**

# DS-2001: Introduction to Data Science

Serial No:

## Final Exam

## Total Time: 3 Hours

## Total Marks: 100

Tuesday, 23rd May, 2023

## Course Instructors

Dr. Ramoza Ahsan

_______________________
Signature of Invigilator

Student Name    Roll No.    Course Section    Student Signature

**DO NOT OPEN THE QUESTION BOOK OR START UNTIL INSTRUCTED.**

**Instructions:**

1. Attempt on question paper. Attempt all of them. Read the question carefully, understand the question, and then attempt it. In case of any ambiguity in the question, state your assumptions and attempt the question.
2. No additional sheet will be provided for rough work. Use the back of the last page for rough work.
3. If you need more space write on the back side of the paper and clearly mark question and part number etc.
4. After asked to commence the exam, please verify that you have **fifteen (15)** different printed pages including this title page. There are total of **6** questions.
5. Calculator sharing is strictly prohibited.
6. Use permanent ink pens only. Any part done using soft pencil will not be marked and cannot be claimed for rechecking.
7. If you have read the instructions, circle course instructor's name to get 5 bonus points.

| | Q-1 | Q-2 | Q-3 | Q-4 | Q-5 | Q-6 | Total |
|---|---|---|---|---|---|---|---|
| **Marks Obtained** | | | | | | | |
| **Total Marks** | 20 | 40 | 10 | 10 | 10 | 10 | 100 |

## Question 1 MCQs [20 Marks]

**Question-1a: What will be the output of the following Python program?**

```
i = 0
while i < 5:
   print(i,end= ‘  ‘)
   i += 1
   if i == 3:
      break
else:
   print(0)
```

a) error

b) 0 1 2 0

**c) 0 1 2**

d) none of the mentioned

**Question-1b: When using the Pandas dropna() method, what argument allows you to change the original DataFrame instead of returning a new one?**

**a) dropna(inplace=True)**

b) dropna(original =True)

c) dropna(keep=True)

d) None of the above

**Question-1c: Which keyword is used for function in Python language?**

a) Function

**b) def**

c) Fun

d) Define

**Question-1d: Which of the following python statement will print the value “7” from the following list**
**A=[1,2,[3,4],[5,6,[7,8]]]**

**a) print(A[3][2][0])**

b) print(A[3][2])

c) print(A[4][2][0]

d) print(A[3][2][1]

**Question-1e: What is the data type of m after the following statement?**
**m = ['July', 'September', 'December']**

a. Dictionary

b. Tuple

**c. List**

d. String

**Question-1f: What is the correct Pandas function for loading CSV files into a DataFrame?**

a. ReadFile()

b. ReadCSV()

c. read_file()

d. **read_csv()**

**Question-1g: What is a correct method to fill empty cells in a data frame with a new value?**

a. value_nul()

b. insertna()

c. replacena()

**d. fillna()**

**Question-1h: Classification is a type of unsupervised machine learning.**

a. True

b. **False**

**Question-1i: Exploratory Data Analysis (EDA) is the analysis of the datasets based on various numerical methods and graphical tools?**

a. **True**

b. False

**Question-1j: Unsupervised learning helps in clustering, association and detection of anomalies in the data.**

a. **True**

b. False

**Question-1k: KNN works well if the number of samples in a dataset are very small.**

a. True

b. **False**

**Question-1l: In the regression equation $y = bo + b_1x$, bo is the;**

a. slope of the line

b. independent variable

**c. y intercept**

d. coefficient of determination

**Question-1m: Which of the following metrics do we have for finding dissimilarity between two clusters in hierarchical clustering?**

1. **Single-link**

2. **Complete-link**

3. **Average-link**

a. 1 and 2

b. 1 and 3

c. 2 and 3

**d. 1, 2 and 3**

**Question-1n: Feature scaling is an important step before applying the K-Mean algorithm. What is the reason behind this?**

a. **In distance calculation, it will give the same weights for all features**

b. You always get the same clusters. If you use or don’t use feature scaling

c. In Manhattan distance, it is an important step, but in Euclidean distance, it is not

**d.** None of these

**Question-1o: Logistic regression is used when you want to:**

a. **Predict a dichotomous variable from continuous or dichotomous variables**.

b. Predict a continuous variable from dichotomous variables.

c. Predict any categorical variable from several other categorical variables.

d. Predict a continuous variable from dichotomous or continuous variables.

**Question-1p: Which of the following steps is performed first by data scientist after acquiring the data?**

**a. Data Cleaning**

b. Data Integration

c. Data Modelling

d. All of the above

**Question-1q: I run a small business, and keep all my business records on an unprotected personal computer. These records include substantial information about my customers. Since I am a small business, I believe I am not a likely target for hackers. Indeed, several years have gone by and no one has stolen any information from my unprotected computer, as far as I know. Are my actions**

**ethical?**

a. Yes

**b. No**

**Question-1r: what are the two kinds of target variables for predictive modeling?**

a. Categorical variable , Nominal Variable

**b. Numerical/Continuous variable, Categorical variable**

c. Numerical/Continuous variable

**d.** Numerical/Continuous variable , ordinal Variable

**Question-1s: How do you handle missing or corrupted data in a dataset?**

a. Drop missing rows or columns

b. Replace missing values with mean/median/mode

c. Assign a unique category to missing values

**d. All of the above**

**Question-1t: Application of Machine learning is _______________**

a. Email filtering

b. Sentiment analysis

c. Face recognition

**d. All of the above**

## Question 2 Short Questions [40 Marks]

**Question-2a [2 points]: Define a dictionary variable “my_dict” with the three keys “Listen”, “Play”, “Study” with the corresponding values “Music”, “Games”, “Hard” and print the value of the key “Play”**

```
my_dict={“Listen”:”Music”, ”Play”:”Games”,”Study”:”Hard”}
print(my_dict[“Play”])
```

**Question-2b [2 points]: List any two supervised machine learning algorithms.**

Decision Trees, KNN, Random Forests, Logistic Regression

**For questions 2c-2f Specify whether the given problem is a supervised or unsupervised machine learning problem. Provide reasoning for your choice.**

**Question-2c [2 points]: You have a dataset of customer transaction records, and you want to group similar customers together for targeted marketing.**

Unsupervised

**Question-2d [2 points]: You are working with a financial institution to detect credit card fraud. The dataset contains a large number of transactions, and only a small percentage of them are fraudulent.**

Supervised

**Question-2e [2 points]: You are tasked with building a recommendation system for an e-commerce platform. The goal is to suggest personalized products to each user based on their browsing and purchase history.**

Unsupervised

**Question-2f [2 points]: You are working on a project to predict the energy consumption of buildings based on various factors, such as size, location, and occupancy.**

Supervised

**Question-2g [4 points]: You are working on a project to predict whether a credit card transaction is fraudulent or legitimate. The dataset you have contains highly imbalanced classes, with a majority of legitimate transactions and only a few fraudulent ones. How would you address the class imbalance issue when training your classification model? Provide a detailed explanation.**

Under sampling, oversampling of data.

**Question-2h [3 points]: In the healthcare domain, you want to predict whether a patient is likely to develop a specific disease based on their medical history. Which machine learning algorithm would you choose for this task, and why? Provide reasoning for your answer!**

This is a classification problem. Decision trees, KNN

**Question-2i [3 points]: You have a dataset with imbalanced class distribution, where the positive class accounts for only 10% of the samples. Additionally, the dataset has a large number of features. You applied a machine learning model to this dataset and obtained the predictions.**

**Which performance metrics would you consider most appropriate for evaluating the model's performance in this scenario? Justify your answer!**

Confusion Matrix, F1 Score

**Question-2j [5 points]: You are developing a regression model to predict the price of houses. The dataset contains a mix of numerical and categorical features, including the neighborhood of each house as a categorical variable. How would you encode the categorical variables to incorporate them into your regression model? Discuss different encoding techniques and their advantages or disadvantages in this scenario.**

Convert the categorical variables into numeric ones by using One Hot Encoding or binary encoding.

**Question-2k [4 points]: Given a confusion matrix for a binary classification problem:**

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| **Actual Positive** | **80** | **20** |
| **Actual Negative** | **10** | **90** |

TP=80, FP=10
FN=20, TN=90

**Compute the following for the model:**

**Accuracy**

TP+TN/(TP+FP+FN+TN)
=(80+90)/(80+10+20+90)
=170/200
=0.85

Or 85%

**F1 Score= 2* Precision * Recall/(Precision + Recall)**
**=2 *0.888*0.8 (0.888+0.8)**
**= 1.4208/(1.688)**
**=0.8417 or 84.2%**

Precision = TP/(TP+FP)
=80/(80+10)
=80/90
=0.888 or 88.8%

Recall = TP/(TP+FN)
=80/(80+20)
=80/100
=0.80 or 80%

**Question-2l [4 points]: Explain the concept of feature selection in machine learning. What are some common techniques used for feature selection, and why is it important?**

A procedure in machine learning to find a subset of features that produces “better” model for a given dataset. Advantages of doing feature selection includes avoid overfitting and achieve better generalization ability, reduce the storage requirements and training time and better interpretability of the model.
Some feature selection/engineering techniques include log transform and scaling. Log transform helps in handling the skewed data, and it makes the distribution more approximate to normal after transformation. With scaling the continuous features become identical in terms of the range. Encoding converts the categorical data in a form so that they can be understood by machine learning algorithms. It enables group of categorical data without losing any information

**Question-2m [2 points]: List any two attribute selection measures that decision tree uses to select an attribute for the node split.**
Information Gain, Gini Index, Gain Ratio

**Question-2n [3 points]: How can we use an unlabeled dataset (without having a target column) in Supervised Learning Algorithms?**
Use clustering to group the datasets and assign each cluster a category value. Use this category value as target value column to train the supervised learning algorithm.

## Question 3 Programming Question [10 Marks]

**Get two input strings (consisting of words and spaces) from the user. Characters can be in lower and upper case. Do the following:**

*a)* Get two input strings from the user and store them in str1 and str2 variables.

```
str1=input("Enter first string").lower()
str2=input("Enter Second string").lower()
```

*b)* For every word in str1, check if it is present in str2, if it is present print it. Also print the count of words in the first string that are present in the second string.

```
str1_words=str1.split()
count=0
for words in str1_words:
    if words in str2:
        print(words)
        count+=1
print("Count of words present in second string are:",count )
```

*c)* Print reverse words of the first string. For example, if str1=”my name is John”, output should be “John is name my”

```
str1=input("Enter first string")
words=str1.split()
```

```
for word in words:
  print(word[::-1])
```

*d)* Swap the first and last word of the two strings and print them. For example if str1=”How are you” and str2=”Have a good trip”, output should be “trip are you”, “Have a good How”

```
str1=input("Enter first string").lower()
str2=input("Enter Second string").lower()
string1=str1.split()
string2=str2.split()
swap=string1[0]
string1[0]=string2[-1]
string2[-1]=swap
print(string1)
print(string2)
```

e) Print a list of all uncommon words between first and second strings. A word is uncommon if it appears exactly once in any one of the strings and does not appear in the other string (case should be ignored). For example, str1=”apple banana mango”, str2=”banana fruits mango”, output=[‘apple’,’fruits’]

**Using dictionary**

```
# count will contain all the word counts

  count = {}
  # insert words of string A to dictionary
  for word in str1.split():
    count[word] = count.get(word, 0) + 1

  # insert words of string str2 to dictionary
  for word in str2.split():
    count[word] = count.get(word, 0) + 1

  # return required list of words
  Print( [word for word in count if count[word] == 1])
```

Using Set

```
A=str1.split()
B=str2.split()
 x=[]
  for i in A:
    if i not in B:
      x.append(i)
  for i in B:
    if i not in A:
      x.append(i)
```

```
print(list(set(x)))
```

## Question 4 Performance Metrics [10 Marks]

**Suppose you are working as an environmental scientist who wants to solve a two-class classification problem for predicting whether a population contains a specific genetic variant. You first tried the Decision Tree model and for the evaluation purposes wants to use a confusion matrix to determine how many ways automated processes might confuse the machine learning classification model. Answer the below questions:**

**Question 4-a [2 points]:** Assume you have 500 samples for your data analysis. In addition, the model predicts that 350 test samples contain the genetic variant, and 150 samples don't. From the dataset you know that the actual number of samples containing the variant is 305, the actual number of samples without the variant is 195, construct the following confusion matrix.

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| **Actual Positive =305** | 305 | 0 |
| **Actual Negative =195** | 45 | 150 |

**Question 4-b [2 points]:** What is TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative) in your scenario.

| | |
|---|---|
| **TP** | **305** |
| **FP** | **45** |
| **TN** | **150** |
| **FN** | **0** |

**Question 4-c [3 points]:** Based on the confusion matrix calculated above, compute the following evaluation metrics.

**Accuracy:**
TP+TN/(TP+FP+FN+TN)= (305+150)/(500)
=0.91 or 91%

**Recall:** = TP/(TP+FN) = 305/(305+0)= 1 or 100%

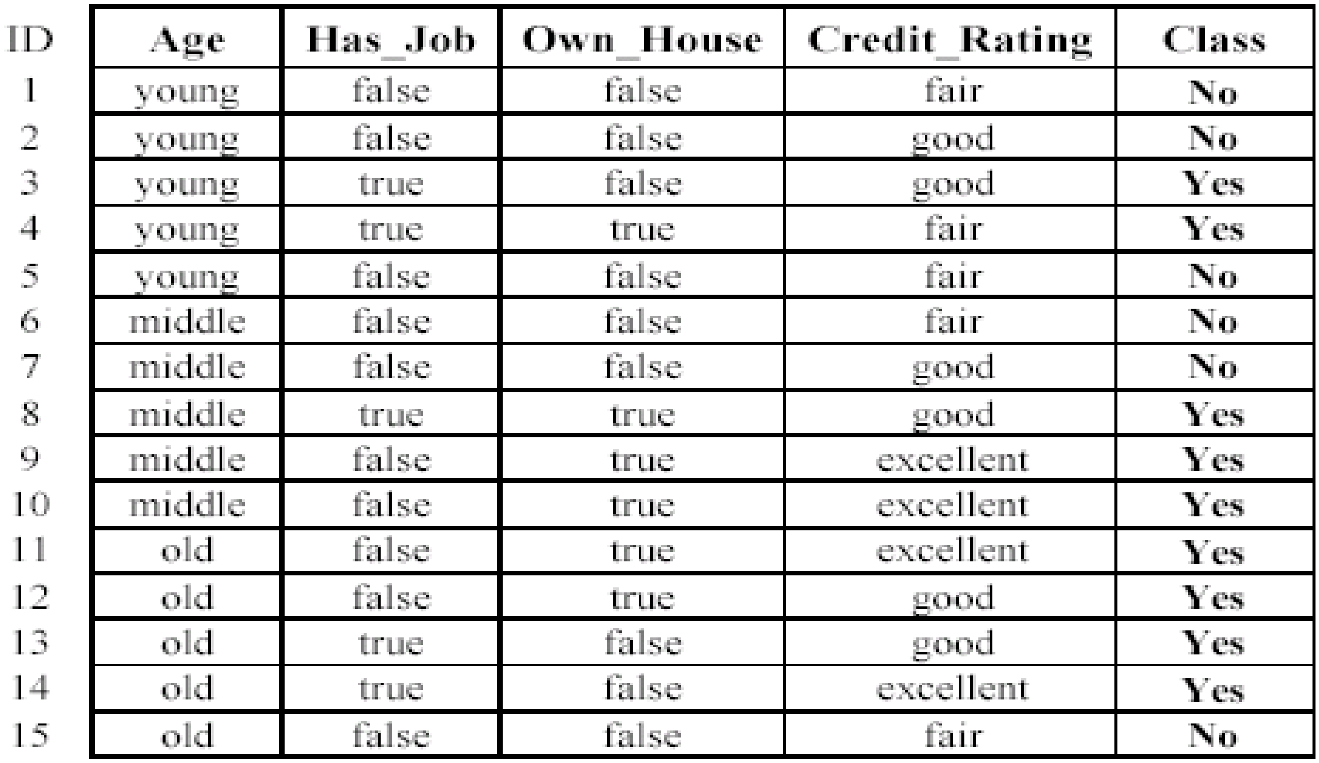
Precision = TP/(TP+FP)= 305/(305+45)= 0.871 or 87.1%

**F1:**
**F1 Score= 2* Precision * Recall/(Precision + Recall)**
**=2*0.871*1/(0.871+1) = 0.931 or 93.1%**

**Question 4-d [3 points]:** Based on the evaluation metrics, what can you say about the model if it is good or bad? Suggest 1 alternate model that you can use for classification task and suggest 1 strategy that you can utilize to increase the performance of your decision tree model.
Based on the evaluation metrics, model is good. We can test out other classification models like logistic regression or KNN. For improving decision tree accuracy can do hyper parameter tuning or prune the tree.

## Question 5 [10 Marks]

**Consider the data set given below that shows whether the loan of the applicant is approved or not. Answer the following questions.**



| ID | Age | Has_Job | Own_House | Credit_Rating | Class |
|---|---|---|---|---|---|
| 1 | young | false | false | fair | No |
| 2 | young | false | false | good | No |
| 3 | young | true | false | good | Yes |
| 4 | young | true | true | fair | Yes |
| 5 | young | false | false | fair | No |
| 6 | middle | false | false | fair | No |
| 7 | middle | false | false | good | No |
| 8 | middle | true | true | good | Yes |
| 9 | middle | false | true | excellent | Yes |
| 10 | middle | false | true | excellent | Yes |
| 11 | old | false | true | excellent | Yes |
| 12 | old | false | true | good | Yes |
| 13 | old | true | false | good | Yes |
| 14 | old | true | false | excellent | Yes |
| 15 | old | false | false | fair | No |

a. **Is the dataset balanced? Provide reasoning for your answer. [2 points]**

No class 6

Yes class 9

Slightly imbalanced

**b. What are the dependent variables and what are the independent variables? [2 points]**

Dependent: Class

Independent: Age, Has_job, Own_house, Credit_Rating

**c. What class label will be assigned to the following unlabeled sample using KNN (K nearest neighbor) assuming k=3? (Show your working. Kindly note the dataset is categorical) [6 points]**

***X={Age=young, Has_Job=true,Own_House=false, Credit_Rating=good, Class=Yes}***

*Class will be yes*

1. 0+1+0+1=2
2. 0+1+0+0=1 - No
3. 0+0+0+0=0 - Yes
4. 0+0+1+1=2
5. 0+1+0+1=2
6. 1+1+0+1=3
7. 1+1+0+0=2
8. 1+0+1+0=2
9. 1+1+1+1=4
10. 1+1+1+1=4
11. 1+1+1+1=4
12. 1+1+1+0=3
13. 1+0+0+0=1 - Yes
14. 1+0+0+1=2
15. 1+1+0+1=3

## Question 6 [10 Marks]

Use **K-means clustering** technique to cluster the following eight points (with (x, y) representing locations) into **three clusters**:

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

$$P(a, b) = |x2 - x1| + |y2 - y1|$$

Use K-Means Algorithm to find the three cluster centers and identify to which cluster each point belongs after the **second** iteration.

First Iteration:

| Given Points | Distance from center (2,10) of Cluster-01 | Distance from center (5,8) of cluster-02 | Distance from center (1,2) of cluster-03 | Point belongs to Cluster |
|---|---|---|---|---|
| A1(2, 10) | 0 | 5 | 9 | C1 |
| A2(2, 5) | 5 | 6 | 4 | C3 |
| A3(8, 4) | 12 | 7 | 9 | C2 |
| A4(5, 8) | 5 | 0 | 10 | C2 |
| A5(7, 5) | 10 | 5 | 9 | C2 |
| A6(6, 4) | 10 | 5 | 7 | C2 |
| A7(1, 2) | 9 | 10 | 0 | C3 |
| A8(4,9) | 3 | 2 | 10 | C2 |

New Cluster Centers

| Cluster-01 | Cluster-02 | Cluster-03 |
|---|---|---|
| A1(2,10) | (6,6) | (1.5,3.5) |

Second Iteration:

| Given Points | Distance from center of Cluster-01 | Distance from center of cluster-02 | Distance from center of cluster-03 | Point belongs to Cluster |
|---|---|---|---|---|
| A1(2, 10) | 0 | 8 | 7 | C1 |
| A2(2, 5) | 5 | 5 | 2 | C3 |
| A3(8, 4) | 12 | 4 | 7 | C2 |
| A4(5, 8) | 5 | 3 | 8 | C2 |
| A5(7, 5) | 10 | 2 | 7 | C2 |
| A6(6, 4) | 10 | 2 | 5 | C2 |
| A7(1, 2) | 9 | 9 | 2 | C3 |
| A8(4,9) | 3 | 5 | 8 | C1 |

New Cluster Centers

| Cluster-01 | Cluster-02 | Cluster-03 |
|---|---|---|
| (3,9.5) | (6.5,5.25) | (1.5,3.5) |

**Best of Luck!**