**1074**

**Code : 20CS51I**

*Register Number*

**V Semester Diploma Examination, June/July-2023**

# ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

**Duration : 3 Hours ]** **[ Max. Marks : 100**

***Instructions :*** Answer **one** full question from each Section.

## SECTION – I

1. (a) Artificial Intelligence (AI) is a promising state-of-the-art technology that provides intelligent solutions in every field today. Justify your answer by describing AI and its applications in various fields. **10**

   (b) Write steps to create repository in GitHub and add file. **10**

2. (a) For the following scenarios you are required to build a predictive model. Which machine learning technique/algorithm can be applied/best suited for stated problems ? **10**

   Justify your recommendations.

   (i) Predicting the food delivery time.

   (ii) Predicting whether the transaction is fraudulent.

   (iii) Predicting the credit limit of a credit card application.

   (iv) Predicting natural disaster.

   (b) How is AI software development life cycle different from traditional software development ? Explain. **10**

## SECTION – II

3. (a) How to handle missing values in the data set ? Explain. **10**

   (b) Perform the following operations on car manufacturing company dataset auto-mpg.csv. Write code given below in Pand/numpy. **10**

   (i) Read data from auto-mpg.csv



[Turn over

(ii) Give code to get all cars with 8 cylinders.

(iii) Get the number of Cars manufactured in each year.

| | mpg | Cylinder | Displace-ment | HP | Weight | Accele-ration | Model year | Origin | Car-name |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 18 | 4 | 300 | 130 | 3504 | 12.0 | 70 | 1 | Honda |
| 1 | 15 | 8 | 325 | 165 | 3695 | 11.5 | 71 | 1 | Nexon |
| 2 | 18 | 8 | 318 | 160 | 3440 | 11.0 | 70 | 1 | Ford |
| 3 | 16 | 6 | 305 | 160 | 3450 | 12.0 | 75 | 1 | Indica |
| 4 | 17 | 8 | 307 | 150 | 3449 | 10.5 | 80 | 1 | Swift |

4. (a) Explain univariate & multivariate data types with examples. **10**

(b) Create two series as shown using pd. series( ) function.

Series – A = [10, 20, 30, 40, 50]

Series – B = [40, 50, 60, 70, 80] **10**

(i) Get the items not common to both.

(ii) Identify the smallest and largest element in the Series A.

(iii) Find the sum of Series B.

(iv) Calculate average in the Series A.

(v) Find median in the given Series B.

## SECTION – III

5. (a) Assume that Iris dataset is given and write the code.

(i) Print first 5 records.

(ii) Print the size of the data for given dataset.

(iii) Use Scatter plot to compare petal length and petal width.

(iv) Check for missing values.

(v) Print the summary of the dataset. **10**

(b) Explain supervised and unsupervised learning with examples. **10**

6. (a) A dataset is given to you for creating a machine learning model. What are the steps followed before using the data for training the model ? Elaborate each followed step. **10**

(b) For the given dataset perform the following operations : **10**

(i) Check Statistical info. of the data set.

(ii) Plot a line chart/plot showing total profit on y-axis and number column on x-axis.

(iii) Find the missing values.

(iv) Find the sum of total profit.

(v) Find the max value from drawing sheets column.

| Number | Pencil | Text Books | Drawing Sheets | Total Units | Profits |
|---|---|---|---|---|---|
| 1 | 300 | 250 | 100 | 800 | 8000 |
| 2 | 350 | 350 | 200 | 1000 | 9500 |
| 3 | 400 | 400 | 200 | 1320 | 10256 |
| 4 | 500 | 420 | 250 | 1510 | 12000 |
| 5 | 520 | 500 | 300 | 2000 | 18000 |

## SECTION – IV

7. (a) The confusion matrix for a machine learning model is given below. Evaluate. **10**

| | | Actual | |
|---|---|---|---|
| | | 1 | 0 |
| Predicted | 1 | 45 | 8 |
| | 0 | 15 | 32 |

(i) Accuracy

(ii) Precision

(iii) Recall

(iv) Specificity

(v) F1-Score

(b) (i) List and briefly explain various activation function in Neural Networks. 5

(ii) Explain neural network Architecture. 5

8. (a) 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 = $(x_1, y_1)$ and b = $(x_2, y_2)$ is defined on :

$$p(a, b) = |x_2 - x_1| + |y_2 - y_1|$$

use K-means Algorithm to find the three clusters after the first iteration. 10

(b) Explain different basic and advanced ensemble learning techniques. 10

## SECTION – V

9. (a) N-grams are defined as combination of N keywords together. Consider the given sentence :

"Machine Learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions relying on patterns and inference instead. It is seen as a subject of artificial intelligence".

(i) Generate unigrams for the above sentence. 5

(ii) Generate trigrams for the above sentence. 5

(b) Summarize public and private cloud deployment models. 10

10. (a) What are ethics in AI and why ethical practices should be followed while developing solutions using AI ? 10

(b) Demonstrate simple linear regression considering a dataset that has two variables :

Salary (dependent variable) and experience (independent variable) 10

**V Semester Diploma Examination, July 2023**

Artificial Intelligence and Machine Learning (20CS51I)

SCHEME OF VALUATION

**Section-I**

**1.** (a) **Any 5 relevant challenges can be considered 2*5=10M**
(b) 10 points or python code **1*10=10M**

**2.** (a) each 2.5 marks 2.5***4=10M**
(b) **any 5 sources 2*5=10M**

**Section-II**

**3.** (a) **List approaches to handle the missing values - 2M + Explanation (2*4=8M)**
(b) (i) 2+(ii)4+(iii)4= 10M

**4.** (a) **explanation with example (5*2=10M)**
**(b) 2 marks for each (2*5=10M)**

**Section-III**

5. (a) 2 marks for each (2*5=10M)
(b) **Explain one method with example** 5M (5*2(methods)=10M)

6. (a) Any five steps(5*2=10M)
(b) **2*5=10M**

**Section IV**

7. (a) for all Equation 1M and Calculation 1M=2M (2*5=10M)
(b) (i) any three activation functions(5M)
(ii) Explain 4M and Example 1M(5M)

8. (a) Calculating the distance from each point to all centres -5M
Finding which point belongs to which cluster and form new clusters 3MCalculate the new cluster centre 2M (5+3+2=10M)
(b) Discuss any five basic and advanced ensemble methods. (2*5=10M)

**Section** V

9. (a) (i) Generate Output of Unigram (5M)
(ii) Generate output of Tri-Gram(5M)

(b) Summarize cloud deployment models . (5*2=10M) 10M

10. (a) Discuss ethical principal and guidelines(5M) in AI with reasons(5M). 10M

(b) Demonstrate Simple Linear Regression considering a dataset that has two variables: salary (dependent variable) and experience (Independent variable)10M
**Importing lib 1M +Reading Dataset 1M + pre-processing 2M+Split, build model 5M +finding score 1M**

## V Semester Diploma Examination, July 2023

Artificial Intelligence and Machine Learning (20CS51I)

### SECTION - I

1. (a)

- Healthcare Advancements: AI has shown great potential in medical diagnosis, drug discovery, and personalized treatment plans. It can analyze medical images, identify patterns of diseases, and assist in surgery, ultimately leading to improved patient outcomes.
- Finance: AI is used for fraud detection, risk management, and portfolio optimization.
- Retail: AI is used for personalizing recommendations, automating customer service, and optimizing supply chain management.
- Transportation: AI is used for traffic prediction, autonomous vehicles, and optimizing logistics.
- Agriculture: AI is used for precision farming, crop monitoring, and weather forecasting.
- Manufacturing: AI is used for predictive maintenance, quality control, and optimizing production processes.
- Education: AI is used for personalizing learning, providing feedback, and automating grading.
- Robotics and Automation: AI has significantly advanced robotics, leading to the development of autonomous vehicles, robotic process automation (RPA), and robotic companions for the elderly and disabled. These applications improve safety, reduce human error, and enhance the quality of life for many.
- Efficiency and Automation: AI systems can process vast amounts of data quickly and accurately, enabling them to automate tasks that would be time-consuming or error-prone for humans. This efficiency leads to increased productivity and cost-effectiveness across industries.
- Data Analysis and Pattern Recognition: AI excels in analyzing complex datasets and identifying patterns that humans may not easily detect. This capability is especially valuable in fields such as finance, healthcare, marketing, and scientific research, where data-driven insights are crucial.
- Personalization: AI-powered algorithms can personalize experiences for individuals, whether in online shopping, content recommendations, or healthcare treatment plans. This level of personalization enhances user satisfaction and engagement.
- Predictive Analytics: AI can predict future trends and outcomes based on historical data, enabling businesses and organizations to make informed decisions and plan for the future effectively.
- Environmental Impact: AI is being used to tackle environmental challenges, such as climate modelling, pollution monitoring, and optimizing energy consumption. It has the potential to contribute significantly to sustainability efforts and create an eco-friendlier world.
- Creativity and Art: AI-generated art, music, and literature are emerging as intriguing fields, where AI can assist and collaborate with human artists, opening new possibilities for creativity.
- Continuous Advancements: The field of AI is rapidly evolving, with ongoing research and development leading to constant improvements. This ensures that AI will continue to push the boundaries of what is possible and drive innovation in various sectors.

(b)

i. To create a repository in GitHub and add a file, follow these steps: Sign in to your GitHub account: If you don't have an account, you'll need to create one at github.com.
ii. Once you're signed in, click on the "+" icon in the top-right corner of the GitHub interface.
iii. Select "New repository": This will take you to the "Create a new repository" page.
iv. Enter a repository name: Choose a descriptive name for your repository. Avoid spaces and special characters, as GitHub will use this name in the repository URL.
v. (Optional) Add a description: Provide a brief description of your repository to help others understand its purpose.
vi. Choose the repository visibility: You can make your repository public, accessible to everyone, or private, accessible only to collaborators you invite.

vii. Initialize with a README file: It's a good practice to initialize the repository with a README file. This file will serve as the home page of your repository, explaining what the project is about.

viii. Add .gitignore and license (optional): You can choose to include a .gitignore file to specify which files or directories should be ignored by version control. Additionally, you can select an open-source license for your project if you wish.

ix. Click on the "Create repository" button: Your new repository will be created, and you'll be redirected to its main page.

x. Clone the repository to your local machine: On the main page of your repository, click on the green "Code" button. Copy the repository URL provided.

- Open a terminal (command prompt) on your local machine, navigate to the directory where you want to store your project, and use the following command to clone the repository: git clone <repository_URL> , Replace <repository_URL> with the URL you copied.
- Add a file to the repository: Create a new file or copy an existing file into the directory you just cloned. For example, let's create a simple file named "example.txt."
- Stage and commit the changes: In your terminal, navigate to the repository's directory and use the following commands: git add example.txt, git commit -m "Added example.txt to the repository"
- Replace "Added example.txt to the repository" with a meaningful commit message describing the changes you made.
- Push the changes to GitHub: To upload your local changes to the remote repository on GitHub, use the following command: git push origin master
- This command will push the changes to the "master" branch. If you want to push to a different branch, replace "master" with the branch name.
- Refresh your GitHub repository page: After pushing the changes, refresh your GitHub repository page, and you should see the "example.txt" file listed.

2. (a) For the following scenarios you are required to build a predictive model. Which machine learning technique/ algorithm can be applied / best suited for stated problems. Justify your recommendation.

- Predicting the food delivery time

ANS: Predicting food delivery time: Regression algorithms such as linear regression or support vector regression would be well-suited for this task as the output is a continuous variable (delivery time) and the goal is to predict a numerical value.

- Predicting whether the transaction is fraudulent.

ANS: Predicting whether a transaction is fraudulent: Classification algorithms such as logistic regression, decision tree, or random forest would be well-suited for this task as the output is a binary variable (fraud or not fraud) and the goal is to predict a class label.

- Predicting the credit limit of a credit card applicant.

ANS: Predicting the credit limit of a credit card applicant: Regression algorithms such as linear regression or support vector regression would be well-suited for this task as the output is a continuous variable (credit limit) and the goal is to predict a numerical value.

Predicting natural disaster.

While AI cannot prevent natural disasters from occurring, it can significantly enhance our ability to predict, monitor, and respond to these events. Here is some ways AI can be applied to predicting natural disasters, Earthquake Prediction, Hurricane and Typhoon Forecasting, Flood Prediction etc.

(b) The development of AI software differs from traditional software development in several key ways. Some of the key differences include:

(i) Data-Driven: AI software development is heavily dependent on data and requires large amounts of high-quality data to train and test models. In traditional software development, data is often an afterthought.

(ii) Experimentation and Iteration: AI software development often involves a lot of experimentation and iteration, as different algorithms and approaches are tried and tested to see which ones work best. Traditional software development is typically more linear and follows a specific plan or design.

(iii) Model Selection: In AI software development, selecting the right model for a particular problem is critical and can be a time-consuming process. In traditional software development, the choice of algorithms and techniques is often predetermined.

(iv) Model evaluation and performance: In AI software development, model performance is evaluated using different metrics and techniques, such as accuracy, precision, recall, and F1 score. In traditional software development, model evaluation is often based on functional requirements.

(v) Deployment and Maintenance: AI software deployment and maintenance requires additional considerations, such as retraining models over time and deploying them in production environments. In traditional software development, deployment and maintenance are often simpler and more straightforward.

(vi) Explainability: AI models are often complex and difficult to understand, which can make it challenging to explain their predictions and decisions to non-technical stakeholders. In traditional software development, the explainability is not as much of an issues.

Overall, AI software development is a more complex, data-driven, and iterative process than traditional software development, requiring specialized knowledge and expertise.

## SECTION – II

3. (a)

There are several ways to handle missing values in a dataset, including:

i. Dropping the rows or columns that contain missing values: This is a simple method, but it can lead to loss of information if the percentage of missing values is high.

Example:

```
df=train_df.drop(['Dependents'],axis=1)
df.isnull().sum()
```

ii. Imputing the missing values: This method involves replacing the missing values with a substitute value, such as the mean or median of the non-missing values. This method can be useful if the percentage of missing values is low, but it can lead to biased results if the missing data is not missing at random.
Example : df =train_df.dropna(axis=0)df.isnull().sum()

iii. Using machine learning algorithms: Some machine learning algorithms can handle missing values automatically and make predictions based on the available data. For example, decision trees and random forests can handle missing values and split the data based on the available features.

iv. Using advanced imputation technique like multiple imputation, this method creates multiple imputed datasets and then combine them using some statistical methods.

- Replacing with Arbitrary Value
  If you can make an educated guess about the missing value, then youcan replace it with some arbitrary value using the following code.
  Ex: In the following code,we are replacing the missing values of the ‘Dependents’column with‘0’train_df['Dependents']=train_df['Dependents'].fillna(0)

- Replacing with Mean
  This is the most common method of imputing missing values of numeric columns.One can use the‘fillna’ method for imputing the columns ‘Loan Amount’ with theme an oft he respective column values as below
  train_df['LoanAmount'].fillna(train_df['LoanAmount'].mean())

- Replacing with Mode
  Mode is the most frequently occurring value. It is used in the case of categorical features.You can use the ‘fillna’ method for imputing the categorical columns ‘Gender’, ‘Married’, and‘Self_Employed’.
  train_df['Gender'].fillna(train_df['Gender'].mode()[0])

- Replacing with Median
  Median is the middlemost value. It’s better to use the median value for imputation in the case

ofoutliers. You can use ‘fillna’ method for imputing the column ‘Loan_Amt’ with the median value.train_df['Loan_Amt']=train_df['Loan_Amt'].fillna(train_df['Loan_Amt'].median()[0])

v. Keep the missing value as is

Sometimes missing data is very less number of rows (say less than 3%) then we cansimply ignore the missing data.There is no hard rule to keep the missing data it depends on us Remove data objects with missing values (Deleting the entire column)

(b)

a) Reading data from an existing file:

```
import pandas as pd
data = pd.read_csv("auto-mpg.csv")
```

b) Get all cars with 8 cylinders:

```
eight_cylinder_cars = data[data['cylinders'] == 8]
print(eight_cylinder_cars)
```

c) Get the number of cars manufactured in each year:

```
cars_by_year = data.groupby('model_year')['name'].count()
print(cars_by_year)
```

4. (a) Univariate Analysis:

- Univariate analysis is the simplest form of statistical analysis that deals with one variable at a time. It is used to describe the basic features of the data in a variable such as the mean, median, mode, and standard deviation.
- For example, a retail store wants to analyse the sales of a particular product. They want to know the average sales, the highest and lowest sales, and the number of products sold during a particular time. This can be done by performing univariate analysis on the sales data.

Multivariate analysis :

- Multivariate analysis is the analysis of more than two variables at a time. It is used to understand the relationship between multiple variables and how they affect each other. It can be used to identify patterns and trends in the data.
- For example, an e-commerce company wants to analyze the relationship between the number of customer reviews, the customer rating, and the number of sales. They can use multivariate analysis to create a plot that shows the relationship between these three variables. This plot can help the company identify patterns and trends in the data, such as which products have the highest sales and which products have the best customer ratings.

(b) Here's one way you could create the two series using the pd.Series() function:

```
import pandas as pd
 Series_A = pd.Series([10,20,30,40,50])
Series_B = pd.Series([40,50,60,70,80])
```

i. To get the items not common to both series A and B, you can use the difference() method:

```
not_common = Series_A.difference(Series_B)
print(not_common)
```

output:

```
 0 10
1 20
2 30
dtype: int64
```

ii. To find the smallest and largest element in the series A, you can use the min() and max() methods:

```
smallest = Series_A.min()
largest = Series_A.max()
```

```
print("Smallest:",smallest)
print("Largest:",largest)
```
output:
Smallest: 10
Largest: 50
iii. To find the sum of series B, you can use the sum() method:
```
sum_B = Series_B.sum()
print(sum_B)
```
output:320
iv. To calculate the average of the series A, you can use the mean() method:
```
average_A = Series_A.mean()
print(average_A)
```
output:30.0
v. To find the median in the series B, you can use the median() method:
```
median_B = Series_B.median()
print(median_B)
```
output:
60.0

**SECTION – III**

5. (a) Here's one way you could work with the Iris dataset using the pandas and matplotlib libraries:

a) Print first 5 records:

```
import pandas as pd

iris = pd.read_csv("iris.csv")

print(iris.head(5))
```

b) Print the size of the data for given dataset

```
print(iris.shape)
```

c) Use scatter plot to compare petal length and petal width

```
import matplotlib.pyplot as plt

plt.scatter(iris['petal_length'], iris['petal_width'])

plt.xlabel('Petal Length')

plt.ylabel('Petal Width')

plt.show()
```

d) Check for missing values:

```
print(iris.isnull().sum())
```

e) Print summarizes of the dataset:

```
print(iris.describe())
```

(b) Supervised Learning and Unsupervised Learning are two fundamental types of machine learning approaches used to train models and make predictions from data.

Supervised Learning: Supervised learning is a type of machine learning where the model is trained on a labeled dataset, meaning the input data is paired with corresponding output labels. The goal of supervised

learning is to learn a mapping function from the input to the output, so the model can make accurate predictions on new, unseen data.

Example: Classification Task: Let's consider a simple example of email classification as "spam" or "not spam" using supervised learning. The dataset contains a collection of emails, each labeled as either "spam" or "not spam," and includes the email's content as the input features.

Unsupervised Learning: Unsupervised learning is a type of machine learning where the model is trained on an unlabeled dataset, meaning the data has no corresponding output labels. The goal of unsupervised learning is to find patterns or structures within the data without explicit guidance.

Example: Clustering Task: Let's consider an example of customer segmentation using unsupervised learning. We have a dataset containing customer transaction data, such as purchase history, spending behavior, and demographics.

6. (a)

(i) Data Exploration: The first step is to explore the data and understand the characteristics of the dataset. This includes understanding the number of observations and variables, the data types of each variable, and the distribution of the data. This can be done by using summary statistics and visualizations such as histograms, box plots, and scatter plots.

(ii) Data Cleaning: The next step is to clean the data. This includes handling missing or corrupted data, removing outliers, and addressing any other data quality issues. This step is important because dirty data can lead to inaccurate or unreliable models.

(iii) Data Transformation: After cleaning the data, it may be necessary to transform the data to make it suitable for the machine learning model. This can include normalizing the data, scaling the data, or creating new variables.

(iv) Feature Selection: Once the data is cleaned and transformed, it is important to select the relevant features that will be used to train the model. This step can be done by using techniques such as correlation analysis, principal component analysis, or mutual information.

(v) Data Splitting: The next step is to split the data into training, validation, and test sets. The training set is used to train the model, the validation set is used to tune the model's parameters, and the test set is used to evaluate the model's performance.

(vi) Feature Engineering: This step is to create new features that will be useful in the model. This can include creating interaction terms, polynomial terms, or binning variables.

(vii) Evaluation Metric: Selecting the right evaluation metric will help to evaluate the model's performance. Common evaluation metrics include accuracy, precision, recall, F1 score, and area under the ROC curve.

(viii) Model Selection: After the data is prepared, the next step is to select the appropriate machine learning model. This can be done by comparing the performance of different models using the evaluation metric.

(ix) Model Training: Once the model is selected, it is trained using the training dataset.

(x) Model Evaluation: Finally, the model's performance is evaluated using the test dataset and the evaluation metric selected.

some of them may be done in parallel or multiple times, depending on the specific dataset and the goal of the analysis.

(b) To perform the given operations on the provided dataset, we'll first organize the data into a tabular format. Here's the data in a table:

| Number | Pencil | Textbooks | Drawing Sheets | Total Units | Profit |
|---|---|---|---|---|---|
| 1 | 300 | 250 | 100 | 800 | 80000 |
| 2 | 350 | 350 | 200 | 1000 | 9500 |
| 3 | 400 | 400 | 200 | 1320 | 10256 |
| 4 | 500 | 420 | 250 | 1510 | 12000 |
| 5 | 520 | 500 | 300 | 2000 | 15000 |

Now, let's perform the requested operations:

i) Check statistical info of the dataset:

We can use Python libraries like Pandas to get statistical information about the dataset.

```
import pandas as pd

data = {
   'Number': [1, 2, 3, 4, 5],
   'Pencil': [300, 350, 400, 500, 520],
   'Textbooks': [250, 350, 400, 420, 500],
   'Drawing Sheets': [100, 200, 200, 250, 300],
   'Total Units': [800, 1000, 1320, 1510, 2000],
   'Profit': [80000, 9500, 10256, 12000, 15000]
}

df = pd.DataFrame(data)
print(df.describe())
```

ii) Plot a line plot showing total profit on the y-axis and the number column on the x-axis:

We can use Python's Matplotlib library to create the line plot.

```
import matplotlib.pyplot as plt

plt.plot(df['Number'], df['Profit'])

plt.xlabel('Number')

plt.ylabel('Total Profit')

plt.title('Total Profit vs. Number')

plt.show()
```

iii) Find the missing values:

To find missing values in the dataset, we can use the isnull() function of Pandas.

```
print(df.isnull().sum())
```

iv) Find the sum of total profit:

```
total_profit_sum = df['Profit'].sum()

print('Sum of Total Profit:', total_profit_sum)
```

v) Find the max value from the Drawing Sheets column:

```
max_drawing_sheets = df['Drawing Sheets'].max()

print('Max value in Drawing Sheets column:', max_drawing_sheets)
```

**SECTION – IV**

7. (a) The confusion matrix looks like this

| | 1 | 0 |
|---|---|---|
| 1 | 45 | 8 |
| 0 | 15 | 32 |

(i) Accuracy:
Accuracy measures the proportion of correctly classified samples out of the total number of samples.
Accuracy = (TP + TN) / (TP + TN + FP + FN) = (45 + 32) / (45 + 8 + 15 + 32) = 77 / 100 ≈ 0.77 or 77%
(ii) Precision:
Precision is the proportion of true positive predictions out of the total positive predictions made by the model.
Precision = TP / (TP + FP) = 45 / (45 + 8) ≈ 0.8491 or 84.91%
(iii) Recall (Sensitivity or True Positive Rate):
Recall measures the proportion of true positive predictions out of all the actual positive samples.
Recall = TP / (TP + FN) = 45 / (45 + 15) ≈ 0.75 or 75%
(iv) Specificity (True Negative Rate):
Specificity measures the proportion of true negative predictions out of all the actual negative samples.
Specificity = TN / (TN + FP) = 32 / (32 + 8) ≈ 0.8 or 80%
(v) F1-Score:
F1-score is the harmonic mean of precision and recall, providing a balanced metric for binary classification when the classes are imbalanced.
F1-Score = 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.8491 * 0.75) / (0.8491 + 0.75) ≈ 0.796 or 79.6%

(b) (i) Activation functions play a crucial role in neural networks by introducing non-linearity to the model's decision-making process. They allow neural networks to learn and approximate complex relationships in data, making them capable of solving a wide range of problems. Below are some popular activation functions used in neural networks:

- Step Function: The step function is one of the simplest activation functions. It maps input values to either 0 or 1 based on a threshold. If the input is greater than or equal to the threshold, it outputs 1; otherwise, it outputs 0. However, the step function is rarely used in modern neural networks due to its lack of differentiability, which is crucial for many optimization algorithms.

- Sigmoid Function (Logistic Function): The sigmoid function maps input values to a range between 0 and 1.

- Hyperbolic Tangent (Tanh) Function: The tanh function is similar to the sigmoid function but maps input values to a range between -1 and 1.
- Rectified Linear Unit (ReLU): ReLU is one of the most popular activation functions used in modern neural networks. It replaces all negative values with zero and leaves positive values unchanged. ReLU is computationally efficient and helps alleviate the vanishing gradient problem. However, it can also suffer from the "dying ReLU" problem, where some neurons get stuck during training and stop learning because they always output zero.
    a. Leaky ReLU:
    b. Parametric ReLU (PReLU)
    c. Exponential Linear Unit (ELU)
    d. Scaled Exponential Linear Unit (SELU)

- Softmax Function: The softmax function is often used in the output layer of a neural network for multi-class classification problems. It converts a vector of real values into a probability distribution, ensuring that the sum of the probabilities of all classes is equal to 1.

(Any three)

(b) (ii) A neural network architecture refers to the structure and organization of the layers, neurons, and connections within a neural network. The architecture of a neural network can vary depending on the problem it is being used to solve, but there are some common building blocks that are used in most neural networks.

The basic building block of a neural network is the artificial neuron, which receives input from other neurons, performs a computation on that input, and generates output. The inputs to a neuron are typically called "features," and the output is typically called the "activation."

A neural network typically consists of an input layer, one or more hidden layers, and an output layer. The input layer receives the raw input data, which is then passed through the hidden layers where computations are performed. The output layer produces the final output of the neural network.

Between the layers, we have weights and biases. Weights are the values that are multiplied with the inputs and the biases are added with the weights. A neural network can be trained by adjusting the weights and biases to minimize the difference between the predicted output and the actual output.
Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are some of the other architectures that are used based on the problem.

8. (a)

Assign each point to the closest initial cluster center. Using the distance function given, we can calculate the distance between each point and each initial cluster center.

- A1(2, 10) is closest to A1(2, 10) with distance = 0
- A2(2, 5) is closest to A1(2, 10) with distance = 5
- A3(8, 4) is closest to A4(5, 8) with distance = 5
- A4(5, 8) is closest to A4(5, 8) with distance = 0
- A5(7, 5) is closest to A4(5, 8) with distance = 3
- A6(6, 4) is closest to A4(5, 8) with distance = 4
- A7(1, 2) is closest to A7(1, 2) with distance = 0
- A8(4, 9) is closest to A4(5, 8) with distance = 1

Recalculate the cluster centers. The new cluster centers are the mean of all the points assigned to each cluster.

- Cluster 1 (A1(2, 10)): {A1(2, 10), A2(2, 5)} - new center = (2, 7.5)
- Cluster 2 (A4(5, 8)): {A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A8(4, 9)} - new center = (6.2, 5.8)
- Cluster 3 (A7(1, 2)): {A7(1, 2)} - new center = (1, 2)
  So the three cluster centers after the first iteration are (2, 7.5), (6.2, 5.8) and (1, 2)

We calculate the distance of each point from each of the centre of the three clusters.The distance is calculated by using the given distance function.
C1 : (2,10) C2 : (5,8) C3 : (1,2)
Calculating distance between A1(2,10) and C1,C2,C3P(A1,C1) = |x2 – x1| + |y2 – y1|
= |2 -2 | + |10 - 10|
= 0
P(A1,C2) = |x2 – x1| + |y2 – y1|
= | 5-2 | + | 8 - 10|
= 5

P(A1, C3) = |x2 – x1| + |y2 – y1|
= | 1-2 | + | 2 - 10|
= 9

Calculating distance between A2(2,5) and C1,C2,C3P(A2,C1) = |x2 – x1| + |y2 – y1|
= |2 -2 | + |10 - 5|
= 5
P(A2,C2) = |x2 – x1| + |y2 – y1|
= | 5-2 | + | 8 - 5|
= 6
P(A2,C3) = |x2 – x1| + |y2 – y1|
= | 1-2 | + | 2 – 5 |
= 4
Calculating distance between A3(8,4) and C1,C2,C3P(A3,C1) = |x2 – x1| + |y2 – y1|
= | 2 -8 | + |10 - 4|
= 12
P(A3,C2) = |x2 – x1| + |y2 – y1|
= | 5-8 | + | 8 – 4 |
= 7
P(A3,C3) = |x2 – x1| + |y2 – y1|
= | 1-8 | + | 2 – 4 |
= 9
Calculating distance between A4(5,8) and C1,C2,C3P(A4, C1) = |x2 – x1| + |y2 – y1|
= | 2 -5 | + |10 -8 |
= 5
P(A4, C2) = |x2 – x1| + |y2 – y1|
= | 5-5 | + | 8 -8 |
= 0
P(A4, C3) = |x2 – x1| + |y2 – y1|
= | 1-5 | + | 2 –8 |
= 10

Calculating distance between A5(7,5) and C1,C2,C3P(A5,C1) = |x2 – x1| + |y2 – y1|
= | 2 - 7 | + |10 - 5 |
= 10
P(A5,C2) = |x2 – x1| + |y2 – y1|
= | 5 - 7 | + | 8 - 5 |
= 5
P(A5,C3) = |x2 – x1| + |y2 – y1|
= | 1 - 7 | + | 2 - 5 |
= 9
Calculating distance between A6(6,4) and C1,C2,C3P(A6,C1) = |x2 – x1| + |y2 – y1|
= | 2 - 6 | + |10 - 4 |
= 10
P(A6,C2) = |x2 – x1| + |y2 – y1|
= | 5 - 6 | + | 8 - 4 |
= 5
P(A6,C3) = |x2 – x1| + |y2 – y1|

= | 1 - 6 | + | 2 - 4 |
= 7
Calculating distance between A7(1,2) and C1,C2,C3P(A7,C1) = |x2 – x1| + |y2 – y1|
= | 2 - 1 | + |10 - 2 |
= 9
P(A7,C2) = |x2 – x1| + |y2 – y1|
= | 5 - 1 | + | 8 - 2 |
= 10
P(A7,C3) = |x2 – x1| + |y2 – y1|
= | 1 - 1 | + | 2 - 2 | =0
Calculating distance between A8(4,9) and C1,C2,C3P(A8,C1) = |x2 – x1| + |y2 – y1|
= | 2 - 4 | + |10 - 9 |
= 3
P(A8,C2) = |x2 – x1| + |y2 – y1|
= | 5 - 4 | + | 8 - 9 |
= 2
P(A8,C3) = |x2 – x1| + |y2 – y1|
= | 1 - 4 | + | 2 - 9 |
= 10

We draw a table showing all the results.Using the table, we decide which point belongs to whichcluster.The given point belongs to that cluster whose centre is nearest to it.

| Given Points | Distance from centre (2, 10) of Cluster-01 | Distance from centre (5, 8) of Cluster-02 | Distance from centre (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 |

From here, New clusters are-

Cluster-01: First cluster contains points-A1(2, 10)
Cluster-02: Second cluster contains points-A3(8, 4),A4(5, 8),A5(7, 5),A6(6, 4),A8(4, 9)Cluster-03: Third cluster contains points-A2(2, 5),A7(1, 2)

- We re-compute the new clusters.
  - The new cluster centre is computed by taking mean of all the points contained in that cluster.For Cluster-01:

We have only one point A1(2, 10) in Cluster-01, so, cluster centre remains the same.For Cluster-02:
Canter of Cluster-02
= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)
= (6, 6)
For Cluster-03:
Canter of Cluster-03
= ((2 + 1)/2, (5 + 2)/2)
= (1.5, 3.5)

This is the completion of Iteration-01.

(b)

(i) Boosting: Boosting is an iterative technique that adjusts the weights of the training instances to focus on difficult cases which increases the accuracy of the base learners. Examples of boosting algorithms are Adaboost, Gradient Boosting, XGBoost etc.

(ii) Bagging: Bagging is a technique that combines multiple models by training them independently on different subsets of the data and averaging their predictions. The Random Forest algorithm is an example of a bagging technique.

(iii) Stacking: Stacking is a technique where multiple models are trained on the same dataset, and the outputs of the models are combined to make the final prediction. The final prediction is made by training a meta-model on the output of the base models.

(iv) Blending: Blending is a technique which is similar to stacking. The only difference is that it uses a small holdout set to train the meta-model, whereas stacking uses the entire dataset.

(v) Hybrid: Hybrid ensemble techniques combine two or more ensemble techniques to improve the performance of the model. For example, combining bagging and boosting.

(vi) Bayesian Model Averaging (BMA): BMA is a method that combines multiple models and assigns a weight to each model based on their performance on a validation set. The final prediction is made by averaging the predictions of all the models, weighted by their assigned Weights.

(vii) Voting (Majority Voting): Voting is a simple and effective ensemble technique where multiple models make predictions, and the final prediction is determined based on the majority vote. In classification tasks, the class with the most votes is chosen as the final prediction. There are two types of voting:

a. Hard Voting: Each model votes for a class label, and the most frequent class label becomes the final prediction.

b. Soft Voting: Each model predicts class probabilities, and the final prediction is obtained by averaging the probabilities across all models and selecting the class with the highest average probability.

(Any Five)

## SECTION – V

9. (a) Text: "Machine learning is the science of developing algorithms and statistical models that computer systems use to perform tasks without explicit instructions, relying on patterns and inference instead. It is seen as a subject of artificial intelligence."

Unigrams:

['machine', 'learning', 'is', 'the', 'science', 'of', 'developing', 'algorithms', 'and', 'statistical', 'models', 'that', 'computer', 'systems', 'use', 'to', 'perform', 'tasks', 'without', 'explicit', 'instructions', 'relying', 'on', 'patterns', 'and', 'inference', 'instead', 'it', 'is', 'seen', 'as', 'a', 'subject', 'of', 'artificial', 'intelligence']

Trigrams:

['machine learning is', 'learning is the', 'is the science', 'the science of', 'science of developing', 'of developing algorithms', 'developing algorithms and', 'algorithms and statistical', 'and statistical models', 'statistical models that', 'models that computer', 'that computer systems', 'computer systems use', 'systems use to', 'use to perform', 'to perform tasks', 'perform tasks without', 'tasks without explicit', 'without explicit instructions', 'explicit instructions relying', 'instructions relying on', 'relying on patterns', 'on patterns and', 'patterns and inference', 'and inference instead', 'inference instead it', 'instead it is', 'it is seen', 'is seen as', 'seen as a', 'as a subject', 'a subject of', 'subject of artificial', 'of artificial intelligence']

(b) 1. Public Cloud

The public cloud makes it possible for anybody to access systems and services. The public cloud may be less secure as it is open to everyone. The public cloud is one in which cloud infrastructure services are provided over the internet to the general people or major industry groups. The infrastructure in this cloud model is owned by the entity that delivers the cloud services, not by the consumer. It is a type of cloud hosting that allows customers and users to easily access systems and services. This form of cloud computing is an excellent example of cloud hosting, in which service providers supply services to a variety of customers. In this arrangement, storage backup and retrieval services are given for free, as a subscription, or on a per-user basis. Example: Google App Engine etc.

Advantages of Public Cloud Model:

- Minimal Investment: Because it is a pay-per-use service, there is no substantial upfront fee, making it excellent for enterprises that require immediate access to resources.
- No setup cost: The entire infrastructure is fully subsidized by the cloud service providers, thus there is no need to set up any hardware.
- Infrastructure Management is not required: Using the public cloud does not necessitate infrastructure management.
- No maintenance: The maintenance work is done by the service provider (Not users).
- Dynamic Scalability: To fulfil company's needs, on-demand resources are accessible.

Disadvantages of Public Cloud Model:

- Less secure: Public cloud is less secure as resources are public so there is no guarantee of high-level security.
- Low customization: It is accessed by many public so it can't be customized according to personal requirements.

2.Private Cloud

The private cloud deployment model is the exact opposite of the public cloud deployment model. It's a one-on-one environment for a single user (customer). There is no need to share your hardware with anyone else. The distinction between private and public clouds is in how you handle all the hardware. It is also called the "internal cloud" & it refers to the ability to access systems and services within a given border or organization. The cloud platform is implemented in a cloud-based secure environment that is protected by powerful firewalls and under the supervision of an organization's IT department. The private cloud gives greater flexibility of control over cloud resources.

Advantages of Private Cloud Model:

- Better Control: You are the sole owner of the property. You gain complete command over service integration, IT operations, policies, and user behaviour.
- Data Security and Privacy: It's suitable for storing corporate information to which only authorized staff have access. By segmenting resources within the same infrastructure, improved access and security can be achieved.
- Supports Legacy Systems: This approach is designed to work with legacy systems that are unable to access the public cloud.
- Customization: Unlike a public cloud deployment, a private cloud allows a company to tailor its solution to meet its specific needs.

Disadvantages of Private Cloud Model:

- Less scalable: Private clouds are scaled within a certain range as there is less number of clients.
- Costly: Private clouds are costlier as they provide personalized facilities.

10. (a) Ethics in AI refers to the principles and guidelines that govern the development and use of AI systems. These principles are designed to ensure that AI systems are developed and used in ways that are fair, transparent, accountable, and respectful of human rights and values.

There are several reasons why ethical practices should be followed while developing solutions using AI. One reason is to prevent harm to individuals or groups of people who may be affected by the AI system. For example, an AI system that is used to make decisions about hiring or lending may perpetuate biases or discrimination if it is not properly designed and tested.

Another reason is to ensure that AI systems are transparent and accountable. This means that the AI system should be able to explain its decision-making process, and that there should be a mechanism for addressing errors or biases in the system. This is important for building trust in the system and ensuring that the people who are affected by the system have a means of redress.

Additionally, Ethical practices in AI can help prevent AI systems from behaving in unexpected or harmful ways. For example, by ensuring that AI systems are designed to be robust and reliable, it can prevent them from behaving erratically or malfunctioning.

Lastly, by respecting human rights and values, Ethical AI can help ensure that AI systems are developed and used in ways that are consistent with human dignity and well-being. This can help ensure that the benefits of AI are shared fairly and that the risks are minimized.

In summary, ethical practices in AI are important to ensure that AI systems are developed and used in ways that are fair, transparent, accountable, and respectful of human rights and values. This can help prevent harm to individuals and groups, build trust in the system, and ensure that AI is used in ways that are consistent with human dignity and well-being.

(b) To implement the Simple Linear regression model in machine learning using Python, we need to follow the below steps:

Step 1: import libraries

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

step 2 :Load the dataset

df = pd.read_csv("salary_data.csv")
Step 3: pre-processing.Check for any missing values and handle it by any suitable method

df.isnull().sum()
mean_A = df['salary'].mean()
df['salary'] = df['salary'].fillna(mean_A)
df.isnull().sum()
Step 4: Split the data set. Extract the dependent and independent variables from the given dataset.

x = df['Experience']
y = df['salary']
x_train, x_test, y_train, y_test = train_test_split(x,y, train_size = 0.8, test_size=0.2, random_state = 21)
x_train=x_train.values.reshape(-1,1)
x_test =x_test.values.reshape(-1,1)
Step 5: Build model

model = LinearRegression()
model.fit(x_train, y_train)
Step 6: Find the accuracy
model.score(x_test,y_test)

Certified that model answers prepared by me for code 20CS51I are from syllabus and scheme of valuation prepared by me is correct.

JAYARAMU H K
Lecturer, Dept. of Comp. Science &Engg.
Govt. Polytechnic Nagamangala-(158)