COMPUTER SCIENCE AND ENGINEERING BOARD SCHEME

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

V Semester Diploma Examination, April/May-2024

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

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

COURSE CODE: 20CS51I

Prepared By:

RAGHAVENDRA M Y

SCHEME OF VALUATION

SECTION-I

1 a)

- Definition of AI: 2M
- Applications of AI in Various Fields (6 marks): 6M
- Critical Analysis and Future Implications (2 marks): 2M

1 a) i) For writing steps to Create a repository in GitHub: 5M

ii) For writing steps to add a file to Repository: 5M

2.a) For each model recommendation 2*4 = 8M

Explanation 2M

2. b) Understanding of Traditional Software Development Life Cycle: 2M

Understanding of AI Software Development Life Cycle: 4M

Comparison of Stages: 4M

Clarity and Coherence: 2M

SECTION-II

3. a) Explanation of any five ways: 10M

b) Listing 2M+Elaboration of each step: 8M

4. a) Each operation from i to v carries 2M each = 5*2M=10M

b) Explanation of each datatype 2*2.5M = 5M + Example for datatype 2*2.5M = 5M

SECTION-III

5. a)

- Writing code: 2 M
- Writing code: 2 M
- Writing code: 2 M
- Writing code: 2 M
- Writing code: 2 M

5. b) Explanation 10M

6.a) i) Importing all required libraries and load data: 2M

ii) Preparing and split data into training and testing data: 3M

iii) Defining model: 3M

iv) Testing model: 2M

6.b) i) Import all required libraries and load data: 2M

ii) Scatter plot to compare area and price: 3M

iii) define model: 3M

iv) test model: 2M

SECTION-IV

7. a) Evaluation of each parameter: 5*2M = 10M

7. b) a) Listing bigrams and trigrams before performing text cleaning steps: 5M

b) i) Listing bigrams and trigrams after performing stop word removal: 2.5M

ii) Listing bigrams and trigrams after replacing punctuations by a single space: 2.5M

8.a) Demonstration with examples: 10M

8.b) Explanation of different techniques: 5M

8.c) Definition: 2M

Explanation of different stages: 3M

SECTION-V

9.a) Diagram: 2M

Explanation of each component: 4*2=8M

9.b) Explaining five V's :1*5M = 5M

9.c) Explaining five challenges: 1*5M = 5M

10.a) Importing lib 1M +Reading Dataset 1M + pre-processing 2M+Split, build model 5M +finding score 1M

10.b) Explanation of each activation function with example: 5* 2 = 10M

NOTE: Any other alternate answer for above any question if found correct/ suitable can also be considered and marks can be awarded accordingly.

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

Raghavendra M Y

Lecturer, Dept of CSE,

Govt Polytechnic Chitradurga

V Semester Diploma Examination

Artificial Intelligence and

Machine Learning (20CS51I)

QUESTION PAPER

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Instructions: Answer one full question from each Section.

SECTION-I

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

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

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? 10M

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. 10M

SECTION-II

3. (a) How to handle missing values in the dataset? Explain. 10M

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

4. a) Create two series as shown using pd.series() function. 10M

Series A = [20, 30, 40, 50, 60]

Series B = [50, 60, 70, 80, 90]

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 mean in the Series A.

v. Find median in the given Series B.

4. b) Referring to the number of variables or features in a dataset and the focus of analysis. Explain univariate & multivariate data types with examples. 10M

SECTION-III

5. (a) Assume that Iris dataset is given and write the code. 10M

- Print first 5 records.
- Print the size of the data for given dataset.
- Use Scatter plot to compare petal length and petal width.
- Check for missing values.
- Print the summary of the dataset.

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

6. a) For the employeesalary.csv dataset perform the following operations

i) Import all required libraries and load data. 2M

ii) prepare and split data into training and testing data 3M

iii) define model 3M

iv) test model 2M

Dataset:

YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445
3.7	57189

6. b) For the houseprices.csv dataset perform the following operations

i) Import all required libraries and load data 2M

ii) Scatter plot to compare area and price 3M

iii) define model 3M

iv) test model 2M

Dataset:

area	price
2600	550000
3000	565000
3200	610000
3600	680000
4000	725000

SECTION-IV

7.a) A machine learning model was built to classify patient as covid +ve or - ve. The confusion matrix for the model is as shown below. Compute other performance metrics and analyse the performance of the model. 10M

		Actual	
		1	0
Predicted	1	397	103
	0	126	142

7.b) N-grams are defined as the combination of N keywords together. Consider the given sentence: 10M

“Data Visualization is a way to express your data in a visual context so that patterns, correlations, trends between the data can be easily understood.” Generate bi-grams and tri-grams for the above sentence

(a)Before performing text cleaning steps.

(b)After performing following text cleaning steps:

(i) Stop word Removal

(ii) Replacing punctuations by a single space

8.a) Demonstrate Stemming and Lemmatization concepts with suitable examples 10M

8.b) Discuss different techniques of cross validation 5M

8.c) What are MLOps? brief different stages that are involved in the MLOps lifecycles 5M

SECTION-V

9.a) With a neat diagram explain components of Docker 10M

9.b) Explain five V's of Big data 5M

9.c) Discuss any five ethical challenges in AI 5M

10.a) Demonstrate simple linear regression considering a dataset that has two variables: Salary (dependent variable) and experience (independent variable) 10M

10.b) Explain activation functions in deep learning using suitable example 10M

MODEL ANSWERS:

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.

1.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.

2. b) The development of AI software differs from traditional software development in several keyways. 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.

3.a) Different approaches to handle the missing values are as follows

1. Keep the missing value as is
2. Remove data objects with missing values (Deleting the entire column)
3. Remove the attributes with missing values (Deleting the entire row)
4. Estimate and impute missing values.

1. Keep the missing value as is sometimes missing data is very less number of rows (say less than 3%) then we can simply ignore the missing data. There is no hard rule to keep the missing data it depends on us.
2. Remove data objects with missing values (Deleting the entire column)

   If a certain column has many missing values, then you can choose to drop the entire column. Code to drop the entire column is as follows:
   df=train_df.drop(['Dependents'],axis=1) df.isnull().sum()

3. Remove the attributes with missing values (Deleting the entire row)
   If a row has many missing values, then you can choose to drop the entire row.

   Code to drop the entire row is as follows:

   df =train_df.dropna(axis=0)

   df.isnull().sum()

4. Estimate and impute missing values

- Replacing with Arbitrary Value
  If you can make an educated guess about the missing value, then you can 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 is the middlemost value. It's better to use the median value for imputation in the case of outliers. 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])

  Note: Code is not mandatory for any approach

3b.

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.

4.a)

Create Series A and Series B series_a = pd.Series([20, 30, 40, 50, 60]) series_b = pd.Series([50, 60, 70, 80, 90])

(i) Get the items not common to both not_common =

series_a[~series_a.isin(series_b)].append(series_b[~series_b.isin(series_a)])

(ii) Identify the smallest and largest element in Series A smallest_: a = series_a.min()

Largest:_a = series_a.max()

(iii) Find the sum of Series B sum_b = series_b.sum()

(iv) Calculate the mean in Series A mean_a = series_a.mean()

(v) Find the median in Series B

median_b = series_b.median()

4.b)

Univariate and multivariate data types refer to the number of variables or features in a dataset and the focus of analysis. These terms are fundamental in statistics and data analysis. Let's explore each type and provide examples:

1. **Univariate Data:** Univariate data analysis deals with a single variable or feature in a dataset. The primary objective is to understand the distribution and characteristics of that individual variable. Univariate analysis is typically used when you want to explore or summarize one aspect of the data.

Example – Univariate Analysis: Suppose you have a dataset containing the ages of a group of people. You are interested in understanding the distribution of ages in this dataset. You perform univariate analysis on the "Age" variable, which may involve creating histograms, calculating summary statistics like mean and median, and visualizing the data using box plots.

2. **Multivariate Data:** Multivariate data analysis involves the analysis of two or more variables simultaneously to understand the relationships and patterns that may exist among them. Multivariate analysis is used when you want to explore how variables interact with each other or when you want to predict one variable based on others. It is common in statistical modeling and machine learning.

Example – Multivariate Analysis: Consider a dataset containing information about houses, including "Square Footage" and "Number of Bedrooms." You are interested in predicting the "Price" of a house based on both of these variables. This scenario involves multivariate analysis, as you are considering the interaction between two variables to predict a third.

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())

5. 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) Import all required libraries and load data

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn. preprocessing import StandardScaler , MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score , mean_squared_error
```

#Reading the data

```
data = pd.read_csv(r"C:\Users\Admin C\Downloads\Salary dataset.csv")
```

ii) prepare and split data into training and testing data

```
X = data.drop("Salary",axis=1)
y = data["Salary"]
X.shape, y.shape
```

Output:

```
((10, 1), (10,))
```

Splitting the data into train and test

```
X_train, X_test, Y_train, Y_test = train_test_split(X,y,random_state=101,test_size=0.2)
X_train.shape, X_test.shape, Y_train.shape, Y_test.shape
```

Output:

((8, 1), (2, 1), (8,), (2,))

iii) define model

```
lr = LinearRegression()
lr.fit(X_train,Y_train)
```

Output:

LinearRegression()

iv) test model

```
pred = lr.predict(X_test)
pred
```

Output:

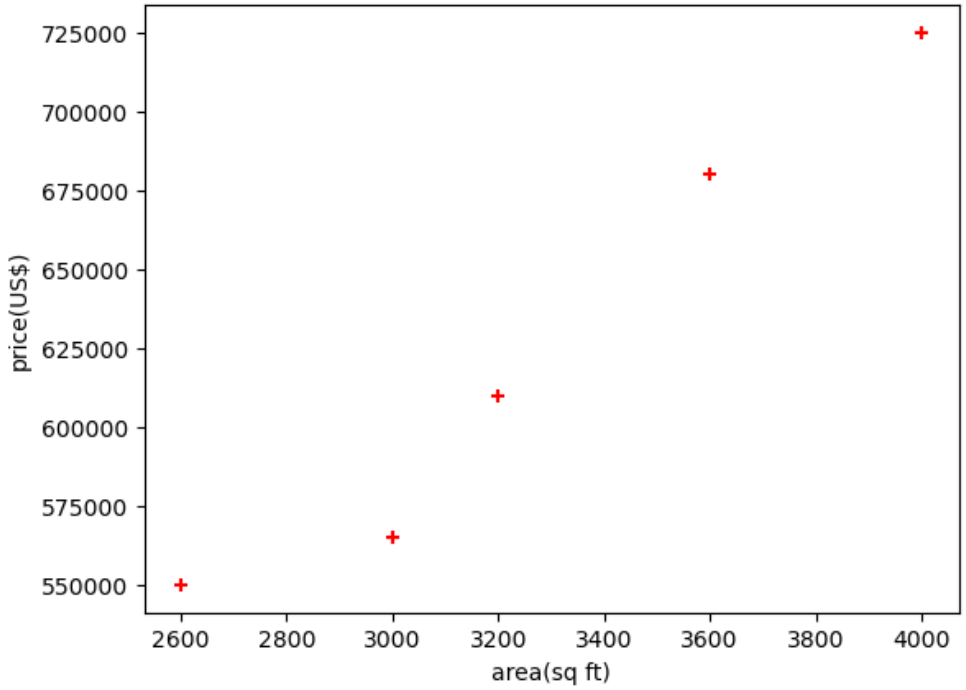
array([55454.8968517 , 42773.67149959])

6.b) i) Import all required libraries and load data

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import linear_model
df = pd.read_csv(r"C:\Users\Admin\ C\Downloads\homeprices.csv")
```

ii) Scatter plot to compare area and price

```
%matplotlib inline
plt.xlabel("area(sq ft)")
plt.ylabel("price(US$)")
plt.scatter(df.area,df.price,color = "red",marker="+")
```



iii) define model

reg = linear_model.LinearRegression()

reg.fit(df[["area"]],df.price)

Output:

LinearRegression()

iv) test the model

reg.predict([[3300]])

Output:

array ([628715.75342466])

SECTION-IV

7.a)

The confusion matrix you provided shows the results of a binary classification model that was trained to predict whether a patient has COVID-19 (positive) or not (negative). The rows represent the actual values, and the columns represent the predicted values. The numbers in the matrix represent the number of observations that fall into each category.

From the confusion matrix, we can compute several performance metrics to evaluate the model's performance:

- **Accuracy**: (397 + 142) / (397 + 103 + 126 + 142) = 0.726 or 72.6%. This metric measures the proportion of correct predictions made by the model.
- **Precision**: 397 / (397 + 103) = 0.793 or 79.3%. This metric measures the proportion of true positive predictions among all positive predictions.
- **Recall (or Sensitivity or True Positive Rate):** 397 / (397 + 126) = 0.760 or 76.0%. This metric measures the proportion of actual positive observations that were correctly predicted as positive.
- **Specificity**: 142 / (142 + 103) = 0.580 or 58.0%. This metric measures the proportion of actual negative observations that were correctly predicted as negative.
- **False Positive Rate (FPR) :** 103 / (142 + 103) = 0.420 or 42.0%. This metric measures the proportion of actual negative observations that were incorrectly predicted as positive.
- **F1-Score:** 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.793 * 0.760) / (0.793 + 0.760) = 0.774 or 77.4%. This is the harmonic mean of precision and recall, which balances both measures.

The model has an accuracy of 72.6%. Which is not a good accuracy. The precision and recall are also not very good. The model is not classifying well.

7.b) Generate bi-grams and tri-grams for the above sentence

a. Before performing text cleaning steps.

ANS: Bi-grams:

- "Data Visualization"
- "Visualization is"
- "is a"
- "a way"
- "way to"
- "to express"
- "express your"
- "your data"
- "data in"
- "in a"
- "a visual"
- "visual context"
- "context so"
- "so that"
- "that patterns"
- "patterns, correlations,"
- "correlations, trends"
- "trends between"
- "between the"
- "the data"
- "data can"
- "can be"
- "be easily"
- "easily understood."

Tri-grams:

❖ "Data Visualization is"
❖ "Visualization is a"
❖ "is a way"
❖ "a way to"
❖ "way to express"
❖ "to express your"
❖ "express your data"
❖ "your data in"
❖ "data in a"
❖ "in a visual"
❖ "a visual context"
❖ "visual context so"
❖ "context so that"
❖ "so that patterns"
❖ "that patterns, correlations,"
❖ "patterns, correlations, trends"
❖ "correlations, trends between"
❖ "trends between the"
❖ "between the data"
❖ "the data can"
❖ "data can be"
❖ "can be easily"
❖ "be easily understood."

b. After performing following text cleaning steps:

i. Stop word Removal:

❖ "Data Visualization"
❖ "Visualization way"
❖ "way express"
❖ "express data"
❖ "data visual"
❖ "visual context"
❖ "context patterns"
❖ "patterns, correlations,"
❖ "correlations, trends"
❖ "trends data"
❖ "data easily"
❖ "easily understood."

Tri-grams:

❖ "Data Visualization way"
❖ "Visualization way express"
❖ "way express data"
❖ "express data visual"
❖ "data visual context"
❖ "visual context patterns"

❖ "context patterns, correlations,"
❖ "patterns, correlations, trends"
❖ "correlations, trends data"
❖ "trends data easily"
❖ "data easily understood."

ii. Replacing punctuations by a single space

ANS: After performing stop word removal and replacing punctuations by

a single space, the sentence becomes:

"Data Visualization way express data visual context patterns correlations

trends data easily understood"

The bi-grams for the cleaned sentence are:

❖ "Data Visualization"
❖ "Visualization way"
❖ "way express"
❖ "express data"
❖ "data visual"
❖ "visual context"
❖ "context patterns"
❖ "patterns correlations"
❖ "correlations trends"
❖ "trends data"
❖ "data easily"
❖ "easily understood"

The tri-grams for the cleaned sentence are:

❖ "Data Visualization way"
❖ "Visualization way express"
❖ "way express data"
❖ "express data visual"
❖ "data visual context"
❖ "visual context patterns"
❖ "context patterns correlations"
❖ "patterns correlations trends"
❖ "correlations trends data"
❖ "trends data easily"
❖ "data easily understood"

8a.

Lemmatization and Stemming are Text Normalization techniques. These techniques are used to prepare words, text, and documents for further processing. **Stemming**

It is the process of producing morphological variants of a root/base word. “boat” would be the stem for [boat, boater, boating, boats].

Lemmatization is a method responsible for grouping different inflected forms of words into the root form, having the same meaning. It is similar to stemming, in turn, it gives the stripped word that has some dictionary meaning.

Lemmatization clearly identifies the base form of ‘troubled’ to ‘trouble’’ denoting some meaning whereas, Stemming will cut out ‘ed’ part and convert it into ‘troubl’ which has the wrong meaning and spelling errors.

‘troubled’ -> Lemmatization -> ‘troubled’, and error

‘troubled’ -> Stemming -> ‘troubl’

Ex for stemming

```
# Import the toolkit and the full Porter Stemmer library
import nltk

from nltk.stem.porter import *

p_stemmer = PorterStemmer()

words = ['run','runner','running','ran','runs','easily','fairly']

for word in words:
    print(word+' --> '+p_stemmer.stem(word))
```

output

```
run --> run
runner --> runner
running --> run
ran --> ran
runs --> run
easily --> easili
fairly --> fairli
```

Ex for lemmatization

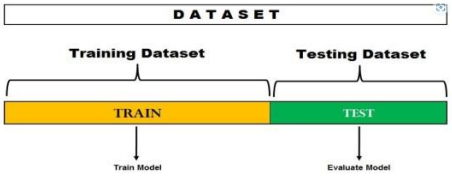
```
import nltk
nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer

# Create WordNetLemmatizer object
wnl = WordNetLemmatizer()

# single word lemmatization examples
list1 = ['kites', 'babies', 'dogs', 'flying', 'plays','feet']
for words in list1:
    print(words + " ---> " + wnl.lemmatize(words))
```

```
kites ---> kite
babies ---> baby
dogs ---> dog
flying ---> flying
plays ---> play
feet ---> foot
```

8b.

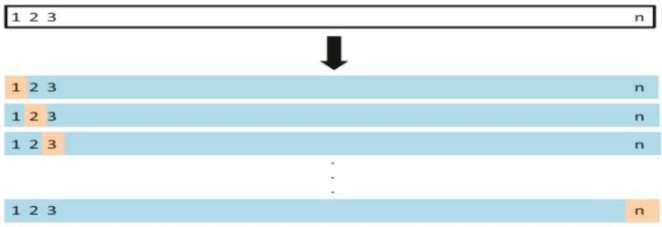
1. Hold Out method

This is the simplest evaluation method and is widely used in Machine Learning projects. Here the entire dataset(population) is divided into 2 sets – train set and test set. The data can be divided into 70-30 or 6040, 75-25 or 80-20, or even 50-50 depending on the use case. As a rule, the proportion of training data has to be larger than the test data.



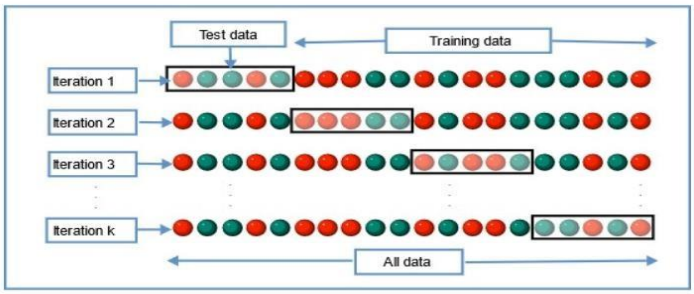
2. Leave One Out Cross-Validation

In this method, we divide the data into train and test sets – but with a twist. Instead of dividing the data into 2 subsets, we select a single observation as test data, and everything else is labelled as training data and the model is trained. Now the 2nd observation is selected as test data and the model is trained on the remaining data.



3. K-Fold Cross-Validation

In this resampling technique, the whole data is divided into k sets of almost equal sizes. The first set is selected as the test set and the model is trained on the remaining k-1 sets. The test error rate is then calculated after fitting the model to the test data.In the second iteration, the 2nd set is selected as a test set and the remaining k-1 sets are used to train the data and the error is calculated. This process continues for all the k sets.



8c.

MLOps stands for Machine Learning Operations. MLOps is focused on streamlining the process of deploying machine learning models to production, and then maintaining and monitoring them. MLOps is a collaborative function, often consisting of data scientists, ML engineers, and DevOps engineers. The word MLOps is a compound of two different fields i.e. machine learning and DevOps from software engineering.

1. ML Development: This is the basic step that involves creating a complete pipeline beginning from data processing to model training and evaluation codes.

2. Model Training: Once the setup is ready, the next logical step is to train the model. Here, continuous training functionality is also needed to adapt to new data or address specific changes.

3. Model Evaluation: Performing inference over the trained model and checking the accuracy/correctness of the output results.

4. Model Deployment: When the proof of concept stage is accomplished, the other part is to deploy the model according to the industry requirements to face the real-life data.

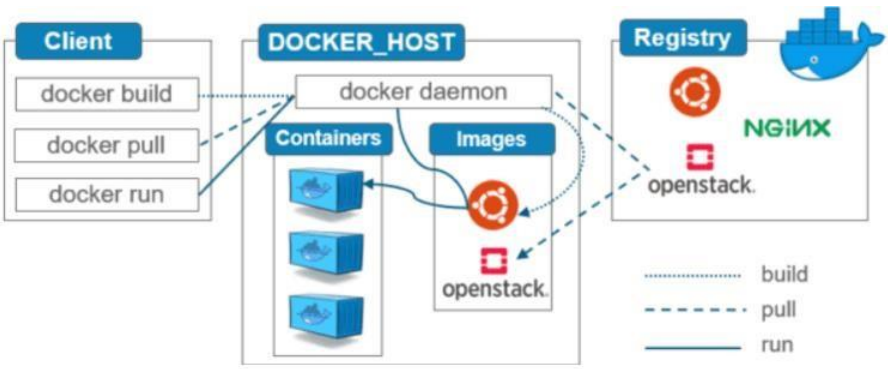
5. Prediction Serving: After deployment, the model is now ready to serve predictions over the incoming data.

6. Model Monitoring: Over time, problems such as concept drift can make the results inaccurate hence continuous monitoring of the model is essential to ensure proper functioning.

7. Data and Model Management: It is a part of the central system that manages the data and models. It includes maintaining storage, keeping track of different versions, ease of accessibility, security, and configuration across various cross-functional teams. Any four stages explanation can be awarded marks

SECTION-V

9a.



Components of Docker

1. Docker Client

Docker client uses **commands** and **REST APIs** to communicate with the Docker Daemon (Server). When a client runs any Docker command on the Docker client terminal, the client terminal sends these Docker commands to the Docker daemon. Docker daemon receives these commands from the Docker client in the form of command and REST API's request.

Docker Client uses Command Line Interface (CLI) to run the following commands -

- Docker build
- Docker pull
- Docker run

2. Docker Registry

Docker Registry manages and stores the Docker images.There are two types of registries in the Docker Pubic Registry - Public Registry is also called as Docker hub**.**

Private Registry - It is used to share images within the enterprise.

3.Docker Daemon: This is the background process that runs on the host machine and manages the containers. It is responsible for creating, starting, stopping, and removing containers, as well as managing their network and storage resources.

4.Docker Images

Docker images are the read-only binary templates used to create Docker Containers. It uses a private container registry to share container images within the enterprise and also uses public container registry to share container images within the whole world. Metadata is also used by docket images to describe the container's abilities.

4.Docker Containers

Containers are the structural units of Docker, which is used to hold the entire package that is needed to run the application. The advantage of containers is that it requires very less resources.In other words, we can say that the image is a template, and the container is a copy of that template.

9.b)

- **Volume:**

  This refers to the amount of data generated every second. With the increase in digital interactions and devices, data accumulates quickly, creating the need for advanced storage and processing solutions.

- **Velocity:**

  This is about the speed at which data is generated, processed, and transformed into insights. Real-time data streams from social media and sensors require quick analysis to harness their potential.

- **Variety:**

  Big Data includes not only numbers but also text, images, videos, and more. Managing and making sense of this diverse data is a challenge. This variety can hold valuable insights but requires flexible tools to process effectively.

- **Veracity:**

  This refers to the reliability and accuracy of the data. Not all data is trustworthy; dealing with inaccuracies is crucial to making informed decisions.

- **Value:**

  The ultimate goal is to extract meaningful insights and Value from the data. This might involve uncovering trends, predicting outcomes, or improving decision-making.

9. c)

- Bias and Discrimination: AI systems can perpetuate and even amplify biases and discrimination if they are not properly designed and tested. For example, an AI system that is used to make decisions about hiring or lending may discriminate against certain groups of people if it is trained on data that contains such biases.
- Privacy and Security: AI systems can collect and process large amounts of personal data, which can raise concerns about privacy and security. For example, an AI system that is used to monitor people's behaviour or predict their behaviour may collect sensitive information about them, which could be used to discriminate against them or to cause harm.
- Lack of Explain ability and Transparency: Many AI systems are based on complex algorithms that are difficult to understand or explain. This can make it difficult for people to understand how the AI system is making its decisions and to hold the system accountable for its actions.
- Job Loss: AI systems can automate many tasks that are currently performed by humans, which can lead to job loss and other economic dislocation. It's important to consider the social and economic impacts of AI and to ensure that the benefits of AI are shared fairly.
- Autonomy and Control: AI systems can be programmed to make decisions and take actions autonomously, which can raise concerns about who is in control of the system and who is responsible for its actions.
- Ethical dilemmas: AI systems may be faced with ethical dilemmas, such as the trade-off between human lives and property damage in self-driving cars, and it may be difficult for the system to make the right decision.
- Societal impact: The development and deployment of AI can have a significant impact on society, and it's important to consider the broader ethical, social, and political implications of AI and to ensure that the technology.

10. a) 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)

10 b. Activation functions in deep learning are mathematical functions applied to the output of each layer in a neural network. They introduce non-linearity into the model, enabling it to learn and represent complex patterns in the data. Without activation functions, a neural network would simply perform linear transformations, limiting its capacity to solve complex problems.

- **Sigmoid Function**

❖ It is a function which is plotted as 'S' shaped graph.
❖ Equation: A = 1/(1 + e-x)
❖ Nature: Non-linear. Notice that X values lies between -2 to 2, Y values are very steep. This means, small changes in x would also bring about large changes in the value of Y.
❖ Value Range: 0 to 1
❖ Example: In binary classification problems, the sigmoid function is often used in the output layer to produce a probability score.
❖ Property: Squashes input values to be between 0 and 1, making it useful for models where the output needs to represent a probability.

- **Tanh Function:**

❖ The activation that works almost always better than sigmoid function is Tanh function also known as **Tangent Hyperbolic function**. It's actually mathematically shifted version of the sigmoid function. Both are similar and can be derived from each other.
❖ **Equation:-**
f(x) = tanh(x) = 2/(1 + e-2x) – 1
OR
tanh(x) = 2 * sigmoid(2x) – 1
❖ **Value Range: -** -1 to +1
❖ **Nature: -** non-linear
❖ **Example:** Tanh is often used in hidden layers of neural networks, especially in natural language processing tasks.

- **RELU Function**

❖ It Stands for Rectified linear unit. It is the most widely used activation function. Chiefly implemented in hidden layers of Neural network.
❖ **Equation: -** A(x) = max(0,x). It gives an output x if x is positive and 0 otherwise.
❖ **Value Range:-** [0, inf)
❖ **Nature: -** non-linear, which means we can easily backpropagate the errors and have multiple layers of neurons being activated by the ReLU function.
❖ **Example:** ReLU is commonly used in hidden layers of convolutional neural networks (CNNs) for tasks like image classification.

- **Softmax Function**

❖ The softmax function is also a type of sigmoid function but is handy when we are trying to handle multi- class classification problems.

- **Nature:** - non-linear
- **Uses: -** Usually used when trying to handle multiple classes. the softmax function was commonly found in the output layer of image classification problems.The softmax function would squeeze the outputs for each class between 0 and 1 and would also divide by the sum of the outputs.
- **Output: -** The softmax function is ideally used in the output layer of the classifier where we are actually trying to attain the probabilities to define the class of each input.
- **Example:** Used in the output layer of a neural network for multi-class classification problems to convert the raw scores into probabilities.