Python

Total Marks: 100

Each question 10 marks

Question 1: -

Write a program that takes a string as input, and counts the frequency of each word in the string, there might be repeated characters in the string. Your task is to find the highest frequency and returns the length of the highest-frequency word.

Note - You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same.

Example input - string = “write write write all the number from from from 1 to 100”

Example output - 5

Explanation - From the given string we can note that the most frequent words are “write” and “from” and the maximum value of both the values is “write” and its corresponding length is 5

Answer 1: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-1.ipynb

Question 2: -

Consider a string to be *valid* if all characters of the string appear the same number of times. It is also *valid* if he can remove just one character at the index in the string, and the remaining characters will occur the same number of times. Given a string, determine if it is *valid*. If so, return `YES` , otherwise return `NO` .

Note - You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same.

Example input 1 - s = “abc”. This is a valid string because frequencies are { “a”: 1, “b”: 1, “c”: 1

} Example output 1- YES

Example input 2 - s “abcc”. This string is not valid as we can remove only 1 occurrence of “c”. That leaves character frequencies of { “a”: 1, “b”: 1 , “c”: 2 }

Example output 2 - NO

Answer 2: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-2.ipynb

Question 3: -

Write a program, which would download the data from the provided link, and then read the data and convert that into properly structured data and return it in Excel format.

Note - Write comments wherever necessary explaining the code written.
Link - https://raw.githubusercontent.com/Biuni/PokemonGO-Pokedex/master/pokedex.json

Data Attributes - id: *Identification Number - int* **num**: *Number of the*

- *Pokémon in the official Pokédex - int* **name**: *Pokémon name -*
- *string* **img**: *URL to an image of this Pokémon - string* **type**:
- *Pokémon type -string* **height**: *Pokémon height - float*

- **weight**: *Pokémon weight - float* **candy**: *type of candy used to evolve Pokémon or given*
- *when transferred - string* **candy_count**: the *amount of candies required to evolve - int*
- **egg**: *Number of kilometers to travel to hatch the egg - float* **spawn_chance**:
- *Percentage of spawn chance* **(NEW) -** float **avg_spawns**: *Number of this pokemon on 10.000 spawns* **(NEW) -** int
- **spawn_time**: *Spawns most active at the time on this field. Spawn times are the same for all time zones and are expressed in local time.* **(NEW) - "**minutes: seconds**" multipliers**: *Multiplier of Combat Power (CP) for calculating the CP after evolution* See below - list of int **weakness**: *Types of*
- *Pokémon this Pokémon is weak to - list of strings* **next_evolution**: *Number and Name of successive evolutions of Pokémon - list of dict* **prev_evolution**: *Number and Name of previous evolutions of Pokémon - - list of dict*

Answer 3: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-3/Python-Answer-3.ipynb

Question 4 -

Write a program to download the data from the link given below and then read the data and convert the into

the proper structure and return it as a CSV file.

Link - https://data.nasa.gov/resource/y77d-th95.json

Note - Write code comments wherever needed for code understanding.
Sample Data -


```
{
        "name": "Tomakovka",
        "id": "24019",
        "nametype": "Valid",
        "recclass": "LL6",
        "mass": "600",
        "fall": "Fell",
        "year": "1905-01-01T00:00:00.000",
        "reclat": "47.850000",
        "reclong": "34.766670",
        "geolocation": {
            "type": "Point",
            "coordinates": [
                34.76667,
                47.85
            ]
        }
    }
```


Excepted Output Data Attributes

- Name of Earth Meteorite - string id - ID of Earth
- Meteorite - int nametype - string recclass - string
- mass - Mass of Earth Meteorite - float year - Year at which Earth
- Meteorite was hit - datetime format reclat - float recclong - float
- point coordinates - list of int

Answer 4: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-4/Python-Answer-4.ipynb

Question 5 -

Write a program to download the data from the given API link and then extract the following data with proper formatting

Link - http://api.tvmaze.com/singlesearch/shows?q=westworld&embed=episodes

Note - Write proper code comments wherever needed for the code understanding
Sample Data -


```
{
            "id": 2326658,
            "url": "https://www.tvmaze.com/episodes/2326658/westworld-4x05-zhuangzi",
            "name": "Zhuangzi",
            "season": 4,
            "number": 5,
            "type": "regular",
            "airdate": "2022-07-24",
            "airtime": "21:00",
            "airstamp": "2022-07-25T01:00:00+00:00",
            "runtime": 60,
            "rating": {
                "average": 7.8
            },
            "image": {
                "medium": "https://static.tvmaze.com/uploads/images/medium_landscape/416/1042460.jpg",
                "original": "https://static.tvmaze.com/uploads/images/original_untouched/416/1042460.jpg"
            },
            "summary": "<p>God is bored.</p>",
            "_links": {
                "self": {
                    "href": "https://api.tvmaze.com/episodes/2326658"
                },
                "show": {
                    "href": "https://api.tvmaze.com/shows/1371"
                }
            }
        }
```


Excepted Output Data Attributes -

- id - int url - string
- name - string season
- - int number - int
- type - string airdate -
- date format airtime -
- 12-hour time format
- runtime - float
- average rating - float
- summary - string
- without html tags
- medium image link - string
- Original image link - string

Answer 5: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-5/Python-Answer-5.ipynb

Question 6 -

Using the data from **Question 3,** write code to analyze the data and answer the following questions **Note**

1. Draw plots to demonstrate the analysis for the following questions for better visualizations.

2. Write code comments wherever required for code understanding

Insights to be drawn -

- Get all Pokemons whose spawn rate is less than 5%
- Get all Pokemons that have less than 4 weaknesses
- Get all Pokemons that have no multipliers at all
- Get all Pokemons that do not have more than 2 evolutions
- Get all Pokemons whose spawn time is less than 300 seconds.

Note - spawn time format is "05:32", so assume "minute: second" format and perform the analysis.

- Get all Pokemon who have more than two types of capabilities

Answer 6: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-6/Python-Answer-6.ipynb

Question 7 -

Using the data from **Question 4**, write code to analyze the data and answer the following questions **Note -** 1. Draw plots to demonstrate the analysis for the following questions for better visualizations

2. Write code comments wherever required for code understanding

Insights to be drawn -

- Get all the Earth meteorites that fell before the year 2000
- Get all the earth meteorites co-ordinates who fell before the year 1970
- Assuming that the mass of the earth meteorites was in kg, get all those whose mass was more than 10000kg

Answer 7: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-7/Python-Answer-7.ipynb

Question 8 -

Using the data from **Question 5,** write code the analyze the data and answer the following questions **Note**

- 1. Draw plots to demonstrate the analysis for the following questions and better visualizations

2. Write code comments wherever required for code understanding

Insights to be drawn -

- Get all the overall ratings for each season and using plots compare the ratings for all the seasons, like season 1 ratings, season 2, and so on.
- Get all the episode names, whose average rating is more than 8 for every season ● Get all the episode names that aired before May 2019
- Get the episode name from each season with the highest and lowest rating
- Get the summary for the most popular (ratings) episode in every season

Answer 8: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-8/Python-Answer-8.ipynb

Question 9 -

Write a program to read the data from the following link, perform data analysis and answer the following questions

Note -

1. Write code comments wherever required for code understanding

Link - https://data.wa.gov/api/views/f6w7-q2d2/rows.csv?accessType=DOWNLOAD

Insights to be drawn -

- Get all the cars and their types that do not qualify for clean alternative fuel vehicle ● Get all TESLA cars with the model year, and model type made in Bothell City.
- Get all the cars that have an electric range of more than 100, and were made after 2015
- Draw plots to show the distribution between city and electric vehicle type

Answer 9: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-9/Python-Answer-9.ipynb

Question 10 -

Write a program to count the number of verbs, nouns, pronouns, and adjectives in a given particular phrase or paragraph, and return their respective count as a dictionary.

Note -

1. Write code comments wherever required for code
2. You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same.

Example Output -

```
dic = {
    "nouns": "count of nouns",
    "pronouns":  "count of pronouns",
    "verbs": "count of verbs",
    "adjectives": "count of adjectives"
}
```

Answer 10: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Python/Python-Answer-10/Python-Answer-10.ipynb

Statistics

Total Marks: 120

Each question 10 marks

Q-1. A university wants to understand the relationship between the SAT scores of its applicants and their college GPA. They collect data on 500 students, including their SAT scores (out of 1600) and their college GPA (on a 4.0 scale). They find that the correlation coefficient between SAT scores and college GPA is 0.7. What does this correlation coefficient indicate about the relationship between SAT scores and college GPA?

Answer 1: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-1/Statistics-Answer-1.ipynb

Q-2. Consider a dataset containing the heights (in centimeters) of 1000 individuals. The mean height is 170 cm with a standard deviation of 10 cm. The dataset is approximately normally distributed, and its skewness is approximately zero. Based on this information, answer the following questions:

a. What percentage of individuals in the dataset have heights between 160 cm and 180 cm?
b. If we randomly select 100 individuals from the dataset, what is the probability that their average height is greater than 175 cm?
c. Assuming the dataset follows a normal distribution, what is the z-score corresponding to a height of 185 cm?
d. We know that 5% of the dataset has heights below a certain value. What is the approximate height corresponding to this threshold?
e. Calculate the coefficient of variation (CV) for the dataset.
f. Calculate the skewness of the dataset and interpret the result.

Answer 2: -
https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-2/Statistics-Answer-2.ipynb

Q-3. Consider the 'Blood Pressure Before' and 'Blood Pressure After' columns from the data and calculate the following

https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_

a. Measure the dispersion in both and interpret the results.
b. Calculate mean and 5% confidence interval and plot it in a graph
c. Calculate the Mean absolute deviation and Standard deviation and interpret the results.
d. Calculate the correlation coefficient and check the significance of it at 1% level of significance.

Answer 3: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-3/Statistics-Answer-3.ipynb

Q-4. A group of 20 friends decide to play a game in which they each write a number between 1 and 20 on a slip of paper and put it into a hat. They then draw one slip of paper

at random. What is the probability that the number on the slip of paper is a perfect square (i.e., 1, 4, 9, or 16)?

Answer 4: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-4/Statistics-Answer-4.ipynb

Q-5. A certain city has two taxi companies: Company A has 80% of the taxis and Company B has 20% of the taxis. Company A's taxis have a 95% success rate for picking up passengers on time, while Company B's taxis have a 90% success rate. If a randomly selected taxi is late, what is the probability that it belongs to Company A?

Answer 5: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-5/Statistics-Answer-5.ipynb

Q-6. A pharmaceutical company is developing a drug that is supposed to reduce blood pressure. They conduct a clinical trial with 100 patients and record their blood pressure before and after taking the drug. The company wants to know if the change in blood pressure follows a normal distribution.

https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_

Answer 6: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-6/Statistics-Answer-6.ipynb

Q-7. The equations of two lines of regression, obtained in a correlation analysis between variables X and Y are as follows:

and . 2�� + 3 − 8 = 0 2�� + ⬚ − 5 = 0 The variance of ⬚ = 4 Find the

a. Variance of Y
b. Coefficient of determination of C and Y
c. Standard error of estimate of X on Y and of Y on X.

Answer 7: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-7/Statistics-Answer-7.ipynb

Q-8. The anxiety levels of 10 participants were measured before and after a new therapy. The scores are not normally distributed. Use the Wilcoxon signed-rank test to test whether the therapy had a significant effect on anxiety levels. The data is given below: Participant Before therapy After therapy Difference

Participant	Before therapy	After therapy	Difference
1	10	7	-3
2	8	6	-2
3	12	10	-2
4	15	12	-3
5	6	5	-1
6	9	8	-1
7	11	9	-2
8	7	6	-1
9	14	12	-2
10	10	8	-2

Answer 8: -
https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-8/Statistics-Answer-8.ipynb

Q-9. Given the score of students in multiple exams

Name	Exam 1	Exam 2	Final Exam
Karan	85	90	92
Deepa	70	80	85
Karthik	90	85	88
Chandan	75	70	75
Jeevan	95	92	96

Test the hypothesis that the mean scores of all the students are the same. If not, name the student with the highest score.

Answer 9: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-9/Statistics-Answer-9.ipynb

Q-10. A factory produces light bulbs, and the probability of a bulb being defective is 0.05. The factory produces a large batch of 500 light bulbs.

a. What is the probability that exactly 20 bulbs are defective?
b. What is the probability that at least 10 bulbs are defective?
c. What is the probability that at max 15 bulbs are defective?
d. On average, how many defective bulbs would you expect in a batch of 500?

Answer 10: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-10/Statistics-Answer-10.ipynb

Q-11. Given the data of a feature contributing to different classes

https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_

a. Check whether the distribution of all the classes are the same or not. b. Check for the equality of variance/
c. Which amount LDA and QDA would perform better on this data for classification and why.
d. Check the equality of mean for between all the classes.

Answer 11: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-11/Statistics-Answer-11.ipynb

Q-12. A pharmaceutical company develops a new drug and wants to compare its effectiveness against a standard drug for treating a particular condition. They conduct a study with two groups: Group A receives the new drug, and Group B receives the standard drug. The company measures the improvement in a specific symptom for both groups after a 4-week treatment period.

a. The company collects data from 30 patients in each group and calculates the mean improvement score and the standard deviation of improvement for each group. The mean improvement score for Group A is 2.5 with a standard deviation of 0.8, while the mean improvement score for Group B is 2.2 with a standard deviation of 0.6. Conduct a t-test to determine if there is a significant difference in the mean improvement scores between the two groups. Use a significance level of 0.05.

b. Based on the t-test results, state whether the null hypothesis should be rejected or not. Provide a conclusion in the context of the study.

Answer 12:-

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Statistics/Statistics-Answer-12/Statistics-Answer-12.ipynb

Machine learning

Total Marks: 210

Each question 15 marks

INTERMEDIATE QUESTIONS :

Q-1. Imagine you have a dataset where you have different Instagram features like u sername , Caption , Hashtag , Followers , Time_Since_posted , and likes , now your task is

to predict the number of likes and Time Since posted and the rest of the features are your input features. Now you have to build a model which can predict the number of likes and Time Since posted.
Dataset This is the Dataset You can use this dataset for this question.

Answer 1: -
https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Machine%20Learning/Machine-Learning-Answer-1/Machine-Learning-Answer-1.ipynb

Q-2. Imagine you have a dataset where you have different features like Age , Gender , Height , Weight , BMI , and Blood Pressure and you have to classify the people into different classes like Normal , Overweight , Obesity , Underweight , and Extreme Obesity by using any 4 different classification algorithms. Now you have to build a model which can classify people into different classes.
Dataset This is the Dataset You can use this dataset for this question.

Answer 2: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Machine%20Learning/Machine-Learning-Answer-2/Machine-Learning-Answer-2.ipynb

Q-3. Imagine you have a dataset where you have different categories of data, Now you need to find the most similar data to the given data by using any 4 different similarity algorithms. Now you have to build a model which can find the most similar data to the given data.
Dataset This is the Dataset You can use this dataset for this question.

Q-4. Imagine you working as a sale manager now you need to predict the Revenue and whether that particular revenue is on the weekend or not and find the Informational_Duration using the Ensemble learning algorithm
Dataset This is the Dataset You can use this dataset for this question.

Answer 4: -

https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Machine%20Learning/Machine-Learning-Answer-4/Machine-Learning-Answer-4.ipynb

Q-5. Uber is a taxi service provider as we know, we need to predict the high booking area using an Unsupervised algorithm and price for the location using a supervised algorithm and use some map function to display the data Dataset
This is the Dataset You can use this dataset for this question.

Q-6. Imagine you have a dataset where you have predicted loan Eligibility using any 4 different classification algorithms. Now you have to build a model which can predict loan Eligibility and you need to find the accuracy of the model and built-in docker and use some library to display that in frontend
Dataset This is the Dataset You can use this dataset for this question.

Answer 6: -
https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Machine%20Learning/Machine-Learning-Answer-6/Machine-Learning-Answer-6.ipynb

Q-7. Imagine you have a dataset where you need to predict the Genres of

Music using

an Unsupervised algorithm and you need to find the accuracy of the model, built-in docker, and use some library to display that in frontend
Dataset This is the Dataset You can use this dataset for this question.

Answer 7: -
https://github.com/seemanshu-shukla/Placement-Assignment_Seemanshu-Shukla/blob/main/Machine%20Learning/Machine-Learning-Answer-7/Machine-Learning-Answer-7.ipynb

Q-8. Quora question pair similarity, you need to find the Similarity between two questions by mapping the words in the questions using TF-IDF, and using a supervised Algorithm you need to find the similarity between the questions. Dataset This is the Dataset You can use this dataset for this question.

Q-9. A cyber security agent wants to check the Microsoft Malware so need he came to you as a Machine learning Engineering with Data, You need to find the Malware using a supervised algorithm and you need to find the accuracy of the model. Dataset This is the Dataset You can use this dataset for this question.

1. An Ad- Agency analyzed a dataset of online ads and used a machine learning model to predict whether a user would click on an ad or not. Dataset This is the Dataset You can use this dataset for this

question. **Advance QUESTIONS :**

Q-1. A Social Media Influencer collected data on Facebook friend requests and used a supervised algorithm to predict whether a user would accept a friend request or not. Dataset This is the Dataset You can use this dataset for this question. Note : Use only Dask and Use MLflow

Q-2. A chemist had two chemical flasks labeled 0 and 1 which consist of two different chemicals. He extracted 3 features from these chemicals in order to distinguish between them, you provided the results derived by the chemicals and your task is to create a model that will label chemical 0 or 1 given its three features and built-in docker and use some library to display that in frontend. Note : Use only pyspark

Dataset This is the Dataset You can use this dataset for this question.

Q- 3. A company wants to predict the sales of its product based on the money spent on different platforms for marketing. They want you to figure out how they can spend money on marketing in the future in such a way that they can increase their profit as much as possible built-in docker and use some library to display that in frontend Dataset This is the Dataset You can use this dataset for this question. Note: Use only Dask

Q-4. Take any 3 questions and deploy them to AWS using GitHub Actions and show a demo link

Q-5. Take any 3 questions and deploy them to AWS using Circle-CI and show a demo link

Deep Learning

Total Marks: 100

Each question 20 marks

Question 1 -

Implement 3 different CNN architectures with a comparison table for the MNSIT dataset using the Tensorflow library.

Note -

1. The model parameters for each architecture should not be more than 8000 parameters
2. Code comments should be given for proper code understanding.
3. The minimum accuracy for each accuracy should be at least 96%

Question 2 -

Implement 5 different CNN architectures with a comparison table for CIFAR 10 dataset using the PyTorch library

Note -

1. The model parameters for each architecture should not be more than 10000 parameters

2 Code comments should be given for proper code understanding

Question 3 -

Train a Pure CNN with less than 10000 trainable parameters using the MNIST Dataset having minimum validation accuracy of 99.40%

Note -

1. Code comments should be given for proper code understanding. 2. Implement in both PyTorch and Tensorflow respectively

Question 4 -

Design an end-to-end solution with diagrams for object detection use cases leveraging AWS cloud services and open-source tech

Note -

1. You need to use both AWS cloud services and open-source tech to design the entire solution

2. The pipeline should consist of a data pipeline, ml pipeline, deployment pipeline, and inference pipeline.
3. In the data pipeline, you would be designing how to get the data from external or existing sources and tech used for the same
4. In the ml pipeline, you would be designing how to train the model, and what all algorithms, techniques, etc. would you be using. Again, tech used for the same 5. Since this is a deep learning project, the use of GPUs, and how effectively are you using them to optimize for cost and training time should also be taken into consideration.
6. In the deployment pipeline, you would be designing how effectively and efficiently you are deploying the model in the cloud,
7. In the inference pipeline, consider the cost of inference and its optimization

related to computing resources and handling external traffic
8. You can use any tool to design the architecture
9. Do mention the pros and cons of your architecture and how much further it can be optimized and its tradeoffs.
10. Do include a retraining approach as well.
11. Try to include managed AWS resources for deep learning like AWS Textract, AWS Sagemaker, etc., and not just general-purpose compute resources like S3, EC2, etc. Try to mix the best of both services

Question 5 -

In **Question 4,** you have designed the architecture for an object detection use case leveraging AWS Cloud, similarly, here you will be designing for Document Classification use case leveraging Azure Cloud services.

Note -

1. Most of the points are the same as in **Question 4,** just cloud services will change

Computer Vision

Total Marks: 200

Each question 20 marks

Question 1 -

Train a deep learning model which would classify the vegetables based on the images provided. The dataset can be accessed from the given link.

Link

https://www.kaggle.com/datasets/misrakahmed/vegetable-image-dataset

Note -

1. Use PyTorch as the framework for training model
2. Use Distributed Parallel Training technique to optimize training time. 3. Achieve an accuracy of at least 85% on the validation dataset.
4. Use albumentations library for image transformation
5. Use TensorBoard logging for visualizing training performance
6. Use custom modular Python scripts to train model
7. Only Jupyter notebooks will not be allowed
8. Write code comments wherever needed for understanding

Question 2 -

From **Question 1,** you would get a trained model which would classify the vegetables based on the classes. You need to convert the trained model to ONNX format and achieve faster inference

Note -

1. There is no set inference time, but try to achieve as low an inference time as possible
2. Create a web app to interact with the model, where the user can upload the image and get predictions
3. Try to reduce the model size considerably so that inference time can be faster
4. Use modular Python scripts to train and infer the model
5. Only Jupyter notebooks will not be allowed
6. Write code comments whenever needed for understanding

Question 3 -

Scrap the images from popular e-commerce websites for various product images sold on those websites. Your goal is to fetch the images from the website, create categories of different product classes and train a deep learning model to classify the same based on the user input.

Note -

1. You can use any framework of your choice like TensorFlow or PyTorch 2. You have to **not use any** pre-trained model, but instead create your own custom architecture and then train the model.
3. Write code comments wherever needed for understanding
4. Try to use little big dataset so that model can be generalized
5. Write modular Python scripts to train and infer the model
6. Only Jupyter Notebook will be not allowed
7. Write code comments wherever needed for code understanding

Question 4 -

You have to train a custom YOLO V7 model on the dataset which is linked below.

Your goal is to detect different products based on the given classes based on the user input

Link -
https://drive.google.com/file/d/1MEgDYJwO_PVVfAbyfjaRHXt7qoiBBHYt/view?usp=share_link

Note -
1. You have to use PyTorch implementation of YOLO V7
2. The dataset consists of 102 classes with train, validation, and test images already in the respective folders.
3. Labeling is already done, given with the dataset, so need for annotation
4. Since the dataset is small, try to achieve at least an mAP of 85 5. Write modular Python scripts to train the model
6. Write code comments wherever needed for understanding
Computer Vision Assessment iNeuron 3
7. Only Jupyter Notebook will not be allowed

Question 5 -
From **Question 4,** you would have a custom-trained YOLO model. Your goal is to need to convert the model to ONNX format and reduce the inference time.

Note -
1. Reduce the inference time to as much as possible
2. Try to reduce the model size by using techniques like Quantization, etc 3. Create a web app for users to interact with your model where users can upload images and get predictions.
4. Write modular Python scripts to infer the model.
5. Only Jupyter notebooks are not allowed.
6. Write code comments wherever needed for code understanding

Question 6 -
You have to train a custom segmentation model based on Detectron 2 framework. Your goal is to segment the given images based on the user input into the different classes

Link -
https://www.kaggle.com/competitions/open-images-2019-instance-segmentation/data

Note -

1. For this, only the Jupyter Notebook is fine
2. Labels are in COCO format.
3. Write code comments wherever needed for understanding

Question 7 -

From **Question 6,** you would have custom trained segmentation model. Your goal is to reduce the model inference time

Note -

1. Reduce inference time to as much as possible
2. Create a web app for users to interact with your model where they can upload images and get predictions
3. Write code comments wherever needed for code understanding.

Question 8 -

You have to train a custom object detection model based on DETR (Detection Transformer)

Link - https://www.kaggle.com/datasets/andrewmvd/helmet-detection

Note -

1. You need to use HuggingFace PyTorch as the framework
2. The dataset is about detecting football players from the images provided
3. Data Annotations are already in COCO format.
4. Write custom Python scripts for training.
5. Write code comments wherever needed for code understanding
6. Only Jupyter Notebooks are not allowed

Question 9 -

From **Question 8,** you would have a custom object detection model **Note -**

1. Try to reduce the model size using quantization
2. Create a web app where the users can interact with your model 3. Write modular Python script for model inference
4. Only Jupyter Notebooks are not allowed
5. Write code comments wherever needed for code understanding

Question 10 -

From all the questions from 1 to 9, take any image classification model, object model detection model, and image segmentation model and deploy it in the cloud **Note -** 1. Deployment of all 3 different models should be AWS, Azure, and GCP 2. A video demo of the application working in the cloud should be good enough 3. Containerization of all 3 applications is important and should be pushed to Docker Hub

4. CI-CD pipelines using GitHub actions that would deploy the models in all 3 clouds are mandatory.

Natural Language Processing

Total Marks: 200

Each question 20 marks

Q-1. Take any YouTube videos link and your task is to extract the comments from that videos and store it in a csv file and then you need define what is most demanding topic in that videos comment section.

Q-2. Take any pdf and your task is to extract the text from that pdf and store it in a csv file and then you need to find the most repeated word in that pdf.

Q-3. from question 2, As you got the CSV and now you need perform key word extraction from that csv file and do the Topic modeling

Q-4. Take any text file and now your task is to Text Summarization without using hugging transformer library

Q-5. Now you need build your own language detection with the fast Text model by Facebook and

Q-6. Generate research papers titles using Bert model and containerize the application and push to public docker hub

Q-7. Now you need to build your own chatbot using the seq2seq model of Amazon website by scrape the website and containerize the application and push to public docker hub

Q-8. Take a any own dataset and build a knowledge bot using Llama model.

Q-9. Using wisher you need transcribe any audio file and then you need to convert that audio file into text file and now convert that text file into audio file of different language.

Q-10. Build a whole End- End api and deploy it on Heroku /railways so the task is that you need build a Auto-Correction of text using NLP
Note: only Jupyter notebook is not allowed from 5th question

← Submission Process →

There are Two Types of Questions Theory based Question and Project-based (where you actually have to code)

First of all, You have to create an Google doc, where you will add answers of all the questions

If you are attempting a question in which you have to write code, so create a repo push your code to repo and copy the link of repo and add it into docs as shown below

Eg. Answer. 6 Python - > GitHub repo link

Note:

- *If you are building any End to end project try to write code in .py file*
- *If you are only analyzing or doing EDA use .ipynb file*

If you are attempting a theory-based question then you have to add the answer in the same google docs as it's

Then submit that final link (google doc link which has all the answers)