**Python**

#### Answer 1: -

https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 2: -

https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 3: -

#### https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 4 -

https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 5 -

https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 6 -

https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 7 -

https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 8 -

https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 9 -

https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

#### Answer 10 –

#### https://github.com/rakku-99/Ineuron\_FSDS\_Assignments/blob/main/Placement%20Assignment\_Rakesh%20Ranjan/PA\_RR.ipynb

# Statistics

**A-1.** A correlation coefficient of 0.7 indicates a strong positive relationship between SAT scores and college GPA.

In this context, a correlation coefficient ranges from -1 to +1, with 0 indicating no relationship, +1 indicating a perfect positive relationship, and -1 indicating a perfect negative relationship.

As SAT scores increase, college GPA tends to increase as well. The correlation coefficient only measures the strength and direction of the linear relationship, but it does not indicate causation or the underlying factors influencing the relationship.

However, it does not provide information about other factors that may contribute to college success, such as study habits, motivation, or personal circumstances.

**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:

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

**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\_

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

**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)?

**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?

**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.

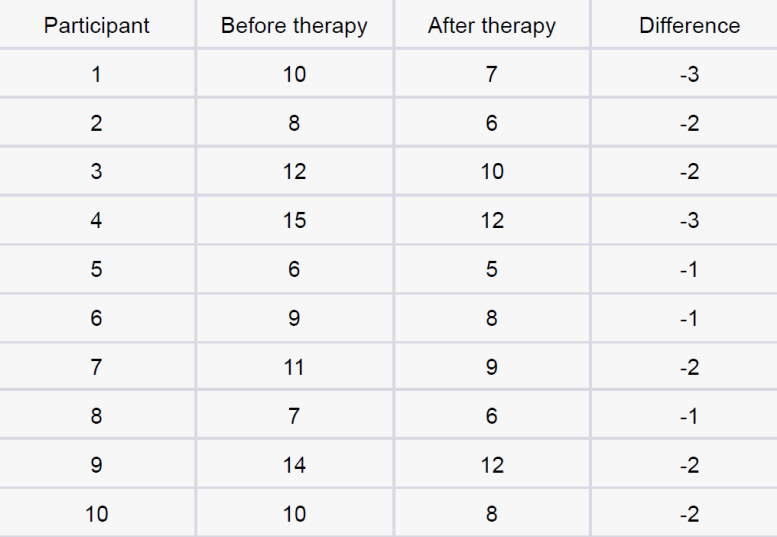
<https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_>

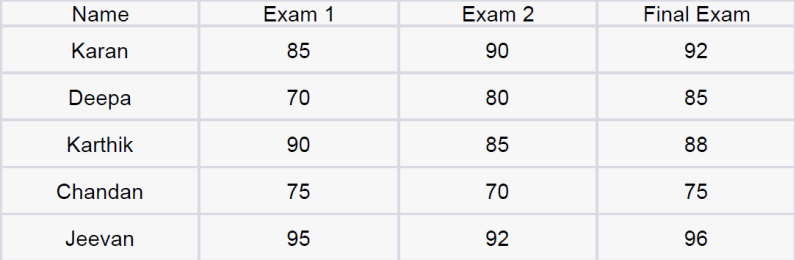
**Q-7.** The equations of two lines of regression, obtained in a correlation analysis between variables X and Y are as follows:

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

1. Variance of Y
2. Coefficient of determination of C and Y
3. Standard error of estimate of X on Y and of Y on X.

**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



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

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

**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.

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

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

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

[=share\_](https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_)

1. Check whether the distribution of all the classes are the same or not.
2. Check for the equality of variance/
3. Which amount LDA and QDA would perform better on this data for classification and why.
4. Check the equality of mean for between all the classes.

**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.

1. 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.
2. 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.

# Machine learning

## Total Marks: 210

**Each Answer 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](https://www.kaggle.com/datasets/rxsraghavagrawal/instagram-reach) This is the Dataset You can use this dataset for this question.

**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](https://www.kaggle.com/datasets/ankurbajaj9/obesity-levels) This is the Dataset You can use this dataset for this question.

**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](https://www.kaggle.com/datasets/rmisra/news-category-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](https://www.kaggle.com/datasets/henrysue/online-shoppers-intention) This is the Dataset You can use this dataset for this question.

**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](https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma) 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](https://www.kaggle.com/code/ajaymanwani/loan-approval-prediction/notebook) This is the Dataset You can use this dataset for this question.

**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](https://www.kaggle.com/datasets/insiyeah/musicfeatures) This is the Dataset You can use this dataset for this question.

**Q-8.** Quora Answer 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](https://www.kaggle.com/c/quora-question-pairs) 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](https://www.kaggle.com/c/microsoft-malware-prediction) 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](https://www.kaggle.com/c/avazu-ctr-prediction) 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. [Datase](https://www.kaggle.com/c/FacebookRecruiting)t 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](https://www.kaggle.com/datasets/uciml/indian-liver-patient-records) 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](https://www.kaggle.com/datasets/ashydv/advertising-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 Answer 20 marks**

#### Answer 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%

#### Answer 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

#### Answer 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

#### Answer 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.

1. In the deployment pipeline, you would be designing how effectively and efficiently you are deploying the model in the cloud,
2. In the inference pipeline, consider the cost of inference and its optimization

related to computing resources and handling external traffic

1. You can use any tool to design the architecture
2. Do mention the pros and cons of your architecture and how much further it can be optimized and its tradeoffs.
3. Do include a retraining approach as well.
4. 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

#### Answer 5 -

In **Answer 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 **Answer 4,** just cloud services will change

# Computer Vision

## Total Marks: 200

**Each Answer 20 marks**

#### Answer 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-

http[s://w](http://www.kaggle.com/datasets/misrakahmed/vegetable-image-dataset)ww.k[aggle.com/datasets/misrakahmed/vegetable-image-dataset](http://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

#### Answer 2 -

From **Answer 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

#### Answer 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.

1. Write code comments wherever needed for understanding
2. Try to use little big dataset so that model can be generalized
3. Write modular Python scripts to train and infer the model
4. Only Jupyter Notebook will be not allowed
5. Write code comments wherever needed for code understanding

#### Answer 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
5. Write code comments wherever needed for understanding

Computer Vision Assessment iNeuron 3

1. Only Jupyter Notebook will not be allowed

#### Answer 5 -

From **Answer 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.
3. Write modular Python scripts to infer the model.
4. Only Jupyter notebooks are not allowed.
5. Write code comments wherever needed for code understanding

#### Answer 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 -

http[s://w](http://www.kaggle.com/competitions/open-images-2019-instance-segmenta)ww.k[aggle.com/competitions/open-images-2019-instance-segmenta](http://www.kaggle.com/competitions/open-images-2019-instance-segmenta) tion/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

#### Answer 7 -

From **Answer 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.

#### Answer 8 -

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

**Link -** http[s://w](http://www.kaggle.com/datasets/andrewmvd/helmet-detection)ww.k[aggle.com/datasets/andrewmvd/helmet-detection](http://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

#### Answer 9 -

From **Answer 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

#### Answer 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

Computer Vision Assessment iNeuron 5

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 Answer 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 Answer 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 Answer and Project-based (where you actually have to code)

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

If you are attempting a Answer 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 Answer 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 Answer s)