ASSIGNMENT FINAL

Submitted By
Name:- Namitha Kolur

PYTHON ASSIGNMENT

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 Python:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_1.ipynb

Additional Test Cases:
1. Input: "hello world hello world hello"
   Output: 5
   Explanation: The most frequent word is "hello" with a frequency of 3. Its length is 5.
2. Input: "apple banana apple orange banana apple"
   Output: 6
   Explanation: The most frequent words are "apple" and "banana" with a frequency of 3. Both words have a length of 6.

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 Python:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_2.ipynb**

Additional Test Case 1:
print(is_valid_string("aabbccd")) # Output: YES

Additional Test Case 2:
print(is_valid_string("aabbccdde")) # Output: NO

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.

Answer 3 Python:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_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 Python:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_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 Python:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_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 Python:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_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 10000 kg

Answer 7 Python:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_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 Python:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_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 Python:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_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.

Answer 10 Python:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Python/Python_10.ipynb

Test Case 1:

The given phrase "The quick brown fox jumps over the lazy dog." contains one verb ("jumps"), one noun ("dog"), and two adjectives ("quick", "lazy"). The counts dictionary will be {'Verbs': 1, 'Nouns': 1, 'Pronouns': 0, 'Adjectives': 2}.

Test Case 2:

The given paragraph contains a mixture of verbs, nouns, pronouns, and adjectives. After counting, the counts dictionary will reflect the respective counts of each category.

Additional Test Case 3:

When an empty string is provided as input, all counts will be zero since there are no words to analyze.

Additional Test Case 4:

In a sentence containing only pronouns, the counts of verbs, nouns, and adjectives will be zero, while the count of pronouns will be equal to the number of words in the sentence (7 in this case).

STATISTICS ASSIGNMENT

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?

Solution:-
A There is a strong correlation between SAT scores and college GPA, with a correlation coefficient of 0.7.

In The positive sign indicates that higher SAT scores are associated with higher college GPAs, and vice versa. This suggests that the SAT scores explain a significant amount of the variability in college GPAs, with a magnitude of 0.7.

It is important to note, however, that correlation does not imply causation. The correlation coefficient suggests a relationship between SAT scores and college GPA, but not that one variable causes the other. A college GPA can also be affected by factors such as study habits, motivation, and external influences.

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.

Solution:-
a. The area under the normal distribution curve between 160 cm and 180 cm can be used to calculate the percentage of individuals with this height. We can detect this area using the Z-table or statistical software because the dataset is roughly normally distributed.

The numbers are first converted to Z-scores using the equation $Z = (X - \mu) / \sigma$, where X is the height, μ is the mean, and σ is the standard deviation.

For 160 cm: $Z1 = (160 - 170) / 10 = -1$
For 180 cm: $Z2 = (180 - 170) / 10 = 1$

To find the region between these Z-scores, we then look up the matching numbers in the Z-table or use statistical tools. The percentage of people who fall within the range of 160 and 180 cm is shown in this region.

Let's assume the area is A.

The dataset's proportion of people with heights between 160 cm and 180 cm is therefore equal to A * 100%.

b. Similar to the original dataset, but with a smaller standard deviation, the average height of 100 randomly chosen people follows a normal distribution. The standard deviation in this situation changes to $\sigma/ \text{sqrt}(n)$, where n is the sample size.

As a result, the standard deviation of the average height for a sample of 100 people is 10 / sqrt(100) = 1 cm.

The Z-score for 175 cm must be calculated using the following formula in order to determine the likelihood that the average height is larger than that measurement: Z is determined by the equation Z = (X -) / (/ sqrt(n)), where X is the value (175 cm), is the mean (170 cm), is the standard deviation (10 cm), and n is the sample size (100).

Z = (175 - 170) / (1) = 5

The probability associated with this Z-score can then be found in the Z-table or by using statistical software.

Let's assume the probability is P.

Therefore, the probability that the average height of a random sample of 100 individuals is greater than 175 cm is P.

c. To find the Z-score corresponding to a height of 185 cm, we use the formula: Z = (X - μ) / σ, where X is the height (185 cm), μ is the mean (170 cm), and σ is the standard deviation (10 cm).

Z = (185 - 170) / 10 = 1.5

Therefore, the Z-score corresponding to a height of 185 cm is 1.5.

d. We need to determine the Z-score corresponding to the cumulative probability of 0.05 in order to determine the approximate height corresponding to the threshold where 5% of the dataset has heights below that value.

Assume that the Z-score is Z.

Using the Z-table or a statistical software, we can find the Z-score corresponding to a cumulative probability of 0.05. Then, we can use the formula: X = Z * σ + μ, where X is the height, Z is the Z-score, σ is the standard deviation (10 cm), and μ is the mean (170 cm).

The height that roughly corresponds to the cutoff point where 5% of the dataset has heights below that amount is therefore X cm.

e. The ratio of the standard deviation to the mean, multiplied by 100% to express it as a percentage, is how the coefficient of variation (CV), a measure of relative variability, is determined.

CV = (σ / μ) * 100%

In this case, CV = (10 / 170) * 100%.

Therefore, the coefficient of variation for the dataset is

CV%.

f. The dataset's about zero skewness suggests that the data is symmetrically distributed. When the skewness is 0, the distribution is fully symmetrical, with identically shaped and sized left and right tails. A symmetrical distribution has an equal mean, median, and mode.

By interpreting this finding, we may conclude that there is no appreciable bias towards higher or lower values and that the heights in the dataset are uniformly distributed around the mean.

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.

Solution:-

a. We can compute the range and interquartile range (IQR) to assess the dispersion in "Blood Pressure Before" and "Blood Pressure After." The IQR denotes the range between the first quartile (Q1) and the third quartile (Q3), whereas the range is the difference between the maximum and minimum values.
For 'Blood Pressure Before':
Range = Maximum value - Minimum value
= 148 - 120
= 28 mmHg
IQR = Q3 - Q1
To find Q1 and Q3:
1. Sort the 'Blood Pressure Before' values in ascending order:
120, 120, 118, 118, 119, 121, 121, 122, 123, 123, 124, 124, 125, 125, 127, 127, 127, 128, 128, 128, 129, 129, 130, 130, 130, 131, 131, 132, 132, 132, 135, 135, 135, 136, 136, 136, 137, 137, 139, 139, 140, 140, 142, 142, 143, 143, 145, 145, 145, 148
2. Calculate Q1 (first quartile):
Q1 = 123
3. Calculate Q3 (third quartile):
Q3 = 136
IQR = Q3 - Q1
= 136 - 123
= 13 mmHg
For 'Blood Pressure After':
Range = Maximum value - Minimum value
= 141 - 118
= 23 mmHg
IQR = Q3 - Q1
To find Q1 and Q3:
1. Sort the 'Blood Pressure After' values in ascending order:
118, 118, 119, 121, 122, 123, 124, 124, 125, 125, 127, 127, 128, 128, 129, 129, 130, 130, 131, 132, 132, 135, 135, 136, 136, 137, 139, 139, 140, 141
2. Calculate Q1 (first quartile):
Q1 = 124

3. Calculate Q3 (third quartile):
Q3 = 136
IQR = Q3 - Q1
= 136 - 124
= 12 mmHg
Interpretation:
Compared to "Blood Pressure After," "Blood Pressure Before" has a higher dispersion. This indicates that the values for "Blood Pressure Before" are more dispersed or diverse than the values for "Blood Pressure After."
b. We must determine the average of the "Blood Pressure Before" and "Blood Pressure After" data in order to determine the mean and 5% confidence interval.
Mean of 'Blood Pressure Before':
(130 + 142 + 120 + 135 + 148 + 122 + 137 + 130 + 142 + 128 + 135 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 +
128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143) / 100
= 132.62 mmHg
Mean of 'Blood Pressure After':
(120 + 135 + 118 + 127 + 140 + 118 + 129 + 124 + 137 + 125 + 129 + 132 + 125 + 136 + 118 + 122 + 130 + 139 + 123 + 132 + 131 + 126 + 120 + 123 + 139 + 122 + 129 + 136 + 131 + 127 + 140 + 119 + 121 + 129 + 137 + 122 + 135 + 131 + 124 + 119 + 124 + 139 + 123 + 131 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137) / 100
= 128.14 mmHg
Confidence Interval:
To calculate the 5% confidence interval, we can use the formula:
Confidence Interval = Mean ± (Critical value * Standard error)
For a 5% confidence level, the critical value is approximately 1.96 (assuming a large sample size).
Standard error = Standard deviation / √n
Standard deviation of 'Blood Pressure Before' (σ_before) = √[Σ(xi - μ_before)^2 / n]
Standard deviation of 'Blood Pressure After' (σ_after) = √[Σ(xi - μ_after)^2 / n]
n = number of observations
Using these formulas, we can calculate the confidence intervals.
c. To calculate the Mean Absolute Deviation (MAD) and Standard Deviation, we need to find the average deviation from the mean for each set of data.
Mean Absolute Deviation (MAD):
MAD_before = Σ|xi - μ_before| / n
MAD_after = Σ|xi - μ_after| / n
Standard Deviation:
Standard deviation_before (σ_before) = √[Σ(xi - μ_before)^2

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)?

Solution:-

To find the probability that the number drawn from the hat is a perfect square, we need to determine the number of favorable outcomes (slips with perfect square numbers) and the total number of possible outcomes (all slips).

The perfect squares between 1 and 20 are 1, 4, 9, and 16.

Favorable outcomes = 4 (since there are 4 perfect square numbers)

Total possible outcomes = 20 (since there are 20 slips in the hat)

Therefore, the probability of drawing a slip with a perfect square number is:

Probability = Favorable outcomes / Total possible outcomes
= 4 / 20
= 0.2

So, the probability of drawing a slip with a perfect square number is 0.2 or 20%.

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?

Solution:-

To solve this problem, we can use Bayes' theorem. Let's define the following events:
- A: The taxi belongs to Company A.
- B: The taxi is late.

We are given:
- P(A) = 0.8 (Company A has 80% of the taxis)
- P(B|A) = 0.05 (Company A's taxis have a 95% success rate, so the probability of being late is 1 - 0.95 = 0.05)
- P(B|not A) = 0.1 (Company B's taxis have a 90% success rate, so the probability of being late is 1 - 0.90 = 0.1)

We want to find P(A|B), which is the probability that the taxi belongs to Company A given that it is late.

Using Bayes' theorem:
P(A|B) = (P(B|A) * P(A)) / P(B)

To calculate P(B), we need to consider the probability of being late regardless of the company:
P(B) = P(B|A) * P(A) + P(B|not A) * P(not A)

P(not A) = 1 - P(A) = 1 - 0.8 = 0.2 (Company B has 20% of the taxis)

Now we can substitute these values into the equation:
P(B) = (0.05 * 0.8) + (0.1 * 0.2) = 0.04 + 0.02 = 0.06

Finally, we can calculate P(A|B):
P(A|B) = (0.05 * 0.8) / 0.06 = 0.04 / 0.06 = 2/3 ≈ 0.6667

Therefore, the probability that a randomly selected taxi is late and belongs to Company A is approximately 0.6667 or 66.67%.

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_

Solution:-

We can run a normality test on the data to see if the change in blood pressure follows a normal distribution. The Shapiro-Wilk test is one that is frequently applied. The test could become extremely sensitive and pick up even little deviations from normalcy with a high sample size (100). We can still administer the test to provide a general evaluation, though.
The Shapiro-Wilk test can be carried out using statistical software or a computer language by following these steps:
1. Enter the data for "Blood Pressure Before" and "Blood Pressure After" into a statistical programme or coding language.
2. For every set of data, perform the Shapiro-Wilk test.
3. For each set of data, find the p-value related to the test.
4. To assess whether the data significantly deviates from a normal distribution, compare the p-values with the significance level (for example, p = 0.05).
We fail to reject the null hypothesis and come to the conclusion that there is no sufficient evidence to imply that the data deviates from a normal distribution if the p-value is higher than the significance level (for example, p > 0.05).
Please take note that before using specific statistical tests or modelling strategies, the normalcy assumptions are frequently checked in real-world circumstances. However, some statistical techniques may be strong enough to produce accurate conclusions even if the data does not exactly match a normal distribution.

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 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

Solution:-

The equations of the regression lines are given as:
2X + 3 - 8 = 0 (1)
2Y + X - 5 = 0 (2)

a. Variance of Y:

We must ascertain the coefficient of X in equation (2) in order to calculate the variance of Y. We can deduce from equation (2) that the coefficient of X is 1. The variance of Y is therefore equal to the variance of X, which is given as 4.
Variance of Y = 4

b. Coefficient of determination (R^2):

The degree to which the regression line fits the data is indicated by the coefficient of determination (R2). It represents the percentage of the dependent variable's (Y) overall fluctuation that can be accounted for by the independent variable (X).

The formula for the coefficient of determination is:

R^2 = (SSR / SST)
where SSR is the sum of squared residuals and SST is the total sum of squares.
To calculate the coefficient of determination, we need to calculate SSR and SST.

SSR is the sum of squared residuals, which can be calculated using the regression equations and the given variance of X:

SSR = Σ(y_predicted - y_actual)^2

Substituting the values into equation (2):

SSR = Σ((2Y + X - 5)^2)

SST is the total sum of squares, which is equal to the variance of Y multiplied by the number of observations:

SST = Variance of Y * n

Substituting the values:

SST = 4 * n

Finally, we can calculate the coefficient of determination:

R^2 = SSR / SST

c. Standard error of estimate:
The average difference between the actual values and the values predicted by the regression line is represented by the standard error of estimation. It serves as a gauge of the regression model's precision. We must figure out the residuals' standard deviation before we can calculate the standard error of estimation for X on Y and Y on X. By taking the square root of the mean squared residual (MSR), the standard deviation of the residuals can be calculated.
Standard error of estimate = √MSR
The sum of squared residuals (SSR), which is equal to the number of observations minus the number of independent variables, must be calculated in order to determine the MSR.

MSR = SSR / df
For X on Y:
df_XonY = n - 1
For Y on X:
df_YonX = n - 2
Substituting the values and calculating the standard error of estimate for both X on Y and Y on X.
Please provide the value of 'n' (number of observations) to proceed with the calculations.

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

Solution:-

Here is the data you provided:

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

To perform the Wilcoxon signed-rank test, follow these steps:

1. Calculate the absolute differences between the before and after therapy scores for each participant.

Participant	Before therapy	After therapy	Difference (Absolute)
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

2. Rank the absolute differences from smallest to largest, ignoring the sign of the differences.

Participant	Difference (Absolute)	Rank
5	1	1
6	1	1
8	1	1
2	2	4
3	2	4
9	2	4
10	2	4
7	2	4
1	3	9
4	3	9

3. Calculate the sum of the positive ranks (W+).
   W+ = 1 + 1 + 1 + 4 + 4 + 4 + 4 + 4 = 23

4. Calculate the sum of the negative ranks (W-).
   W- = 9 + 9 = 18

5. Determine the smaller of W+ and W- (T).
T = min(W+, W-) = min(23, 18) = 18

6. Calculate the expected value of T under the null hypothesis of no difference (E(T)).
E(T) = (n(n + 1)) / 4 = (10(10 + 1)) / 4 = 27.5

7. Calculate the standard deviation of T under the null hypothesis (SD(T)).
SD(T) = sqrt((n(n + 1)(2n + 1)) /

24) = sqrt((10(10 + 1)(2(10) + 1)) / 24) = sqrt(385 / 24) ≈ 3.49

8. Calculate the standardized test statistic (Z).
Z = (T − E(T)) / SD(T) = (18 − 27.5) / 3.49 ≈ −2.73

9. Look up the critical value for a two-tailed test with 10 participants and a significance level (α) of 0.05. The critical value is -1.96.

10. Compare the calculated Z value with the critical value:
- If the calculated Z value is greater than the critical value, reject the null hypothesis.
- If the calculated Z value is less than the negative of the critical value, reject the null hypothesis.
- Otherwise, fail to reject the null hypothesis.

In this case, -2.73 is less than -1.96, so we reject the null hypothesis. This indicates that the therapy had a significant effect on anxiety levels.

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.

Solution:-

A one-way analysis of variance (ANOVA) test can be used to verify the claim that all students' mean scores are equal. The alternative hypothesis (Ha) states that at least one mean score differs from the others, while the null hypothesis (H0) states that the mean scores of all the students are equal.

Here are the scores of the students in the three 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

Let's calculate the mean score for each student and perform the ANOVA test:

Step 1: Calculate the mean score for each student.

Name	Mean Score
Karan	89
Deepa	78.33
Karthik	87.67
Chandan	73.33
Jeevan	94.33

Step 2: Calculate the overall mean score (grand mean).

Grand Mean = (89 + 78.33 + 87.67 + 73.33 + 94.33) / 5 = 84.133

Step 3: Calculate the sum of squares within groups (SSW).

$SSW = \Sigma (Xi - \bar{X}i)^2$

For each student:
Karan: $(85 - 89)^2 + (90 - 89)^2 + (92 - 89)^2 = 18$
Deepa: $(70 - 78.33)^2 + (80 - 78.33)^2 + (85 - 78.33)^2 = 80.33$
Karthik: $(90 - 87.67)^2 + (85 - 87.67)^2 + (88 - 87.67)^2 = 4.33$
Chandan: $(75 - 73.33)^2 + (70 - 73.33)^2 + (75 - 73.33)^2 = 6.67$
Jeevan: $(95 - 94.33)^2 + (92 - 94.33)^2 + (96 - 94.33)^2 = 4.67$

SSW = 18 + 80.33 + 4.33 + 6.67 + 4.67 = 114

Step 4: Calculate the sum of squares between groups (SSB).

$SSB = \Sigma Ni * (\bar{X}i - \bar{X})^2$

For each student:

Karan: 3 * (89 - 84.133)^2 ≈ 57.296
Deepa: 3 * (78.33 - 84.133)^2 ≈ 65.772
Karthik: 3 * (87.67 - 84.133)^2 ≈ 11.925
Chandan: 3 * (73.33 - 84.133)^2 ≈ 107.659
Jeevan: 3 * (94.33 - 84.133)^2 ≈ 305.859

SSB = 57.296 + 65.772 + 11.925 + 107.659 + 305.859 ≈ 548.511

Step 5: Calculate the degrees of freedom (df).

df between = k - 1 =

5 - 1 = 4 (k is the number of groups)
df within = N - k = 15 - 5 = 10 (N is the total number of observations)

Step 6: Calculate the mean square between (MSB) and the mean square within (MSW).

MSB = SSB / df between = 548.511 / 4 ≈ 137.128
MSW = SSW / df within = 114 / 10 ≈ 11.4

Step 7: Calculate the F-statistic.

F = MSB / MSW = 137.128 / 11.4 ≈ 12.019

Step 8: Look up the critical F-value for a significance level (α) of your choice and with df between = 4 and df within = 10. Let's assume α = 0.05.

For α = 0.05, the critical F-value is approximately 3.10.

Step 9: Compare the calculated F-statistic with the critical F-value.

If the calculated F-statistic is greater than the critical F-value, reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

In this case, the calculated F-statistic (12.019) is greater than the critical F-value (3.10). Therefore, we reject the null hypothesis.

Conclusion: Based on the ANOVA test, we can conclude that the mean scores of the students are not the same.

To determine the student with the highest score, we can compare their mean scores. In this case, Jeevan has the highest mean score of 94.33.

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?

Solution:-

To solve these probability questions, we will use the binomial probability formula:

P(X=k) = C(n, k) * p^k * (1-p)^(n-k)

Where:
- P(X=k) is the probability of getting exactly k successes (defective bulbs)
- n is the total number of trials (total number of bulbs)
- k is the number of successes (number of defective bulbs)
- p is the probability of success (probability of a bulb being defective)
- (1-p) is the probability of failure (probability of a bulb not being defective)
- C(n, k) is the number of combinations, calculated as C(n, k) = n! / (k! * (n-k)!)

a. What is the probability that exactly 20 bulbs are defective?

P(X=20) = C(500, 20) * (0.05^20) * (0.95^(500-20))

b. What is the probability that at least 10 bulbs are defective?

P(X >= 10) = P(X=10) + P(X=11) + ... + P(X=500)

c. What is the probability that at most 15 bulbs are defective?

P(X <= 15) = P(X=0) + P(X=1) + ... + P(X=15)

d. On average, how many defective bulbs would you expect in a batch of 500?

The expected value of a binomial distribution is given by E(X) = n * p. So, the expected number of defective bulbs would be E(X) = 500 * 0.05.

To get the specific probabilities and expected value, we can use a statistical software, such as Excel or Python, or use statistical tables.

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.

Solution:-

a. We can use analysis of variance (ANOVA) to see if the distribution of all the classes is the same or not. ANOVA examines whether the means of two or more groups differ statistically significantly from one another. In this instance, we may contrast the blood pressure readings for each class before and after treatment.

b. We can run a test like Bartlett's test or Levene's test to see if the variances are equal. These tests determine whether there are statistically significant differences in the variances of various groups. In this instance, we may contrast the variations between each class's blood pressure readings before and after treatment.

c. When predicting the class membership of observations based on their predictor factors, classification methods such as LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis) are used. LDA makes the assumption that all classes have identical covariance matrices and only differ in terms of means. Contrarily, QDA permits various covariance matrices for every class.
We can run cross-validation or assess the misclassification rate for both approaches to evaluate which method (LDA or QDA) would perform better on this data for classification. Which method offers a higher level of classification accuracy can be determined by contrasting the results of LDA and QDA on the data. The underlying data distribution and the covariance matrices' underlying assumptions determine whether to use LDA or QDA.

d. One-way ANOVA or paired t-tests can be used to determine whether the mean is the same across all classes. These analyses can establish whether there are statistically significant variations between the means of various classes. We can determine if there are appreciable differences in means between the classes by comparing the means of the blood pressure before and after values for each class.

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.

Solution:-

a. A two-sample independent t-test can be used to assess whether there is a significant difference in the mean improvement scores between the two groups (Group A and Group B). The alternative hypothesis (Ha) states that there is a significant difference, contrary to the null hypothesis (H0), which states that there is no significant difference in the mean improvement scores between the two groups.

Given:

Group A: Mean improvement score (μA) = 2.5, Standard deviation (σA) = 0.8, Sample size (nA) = 30
Group B: Mean improvement score (μB) = 2.2, Standard deviation (σB) = 0.6, Sample size (nB) = 30

Using the formula for the two-sample independent t-test:

t = (μA - μB) / sqrt((σA^2/nA) + (σB^2/nB))

Substituting the given values:

t = (2.5 - 2.2) / sqrt((0.8^2/30) + (0.6^2/30))

Calculating the value of t:

t = 0.3 / sqrt((0.0173) + (0.012))

b. The critical value for a significance level of 0.05 with (nA + nB - 2) degrees of freedom is obtained from a t-distribution table. Let's assume it to be tcrit.

The null hypothesis (H0) is rejected and we come to the conclusion that there is a substantial difference in the mean improvement scores between the two groups if the estimated value of t is greater than tcrit or falls in the critical zone (t > tcrit). In the absence of this, we fail to reject the null hypothesis and come to the conclusion that there is no significant difference if the computed value of t is smaller than tcrit (t tcrit).

Without knowing the crucial value or degrees of freedom, we cannot predict the outcome of the t-test because we don't have the exact values. We can, nevertheless, conclude that there is evidence to support a substantial difference in the mean improvement scores between Groups A and B based on the available data and the assumption that the estimated t-value is bigger than the critical value.

MACHINE LEARNING ASSIGNMENT

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

Answer 1 ML:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Machine_Learning/ML_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 ML:-

https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Machine_Learning/ML_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. Link:- https://www.kaggle.com/datasets/rmisra/news-category-dataset/download?datasetVersionNumber=3
This is the Dataset You can use this dataset for this question.

Answer 3 ML:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Machine_Learning/ML_3.ipynb

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 ML:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Machine_Learning/ML_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.

Answer 5 ML:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Machine_Learning/ML_5.ipynb

DEEP LEARNING ASSIGNMENT

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%

Answer 1 DL:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Deep_Learning/DL_1.ipynb

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.

Answer 2 DL:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Deep_Learning/DL_2.ipynb

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.

Answer 3 DL:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Deep_Learning/DL_3.ipynb

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.

Answer 4 DL:-

We can develop a comprehensive architecture with a data pipeline, ML pipeline, deployment pipeline, and inference pipeline to create an end-to-end solution for object detection use cases using AWS cloud services and open-source technology. An overview of the solution architecture is provided below:

Let's look over each pipeline element and the employed technologies:

1. Data Pipeline:
- External or existing data sources: These can be datasets from a variety of sources, such as open datasets, datasets with custom labels, or data produced by Internet of Things (IoT) devices.
- Data storage: To store the input data in a scalable and long-lasting way, use AWS S3. High availability, durability, and support for integrations with other AWS services are all features of S3.
- Data preprocessing: Perform data augmentation, resizing, normalisation, and other essential preprocessing procedures using open-source technology such as TensorFlow or OpenCV.
- Data annotation: To manually identify the objects in the photos and provide bounding box annotations, use open-source annotation tools like LabelImg or RectLabel.
- Annotation storage: To make it simple to access the annotations during training, store them in a structured format, like JSON or XML, in S3 or a database.

2. ML Pipeline:
- Training infrastructure: Make use of AWS SageMaker. Managed Jupyter notebooks, distributed training capabilities, and support for well-liked deep learning frameworks like TensorFlow and PyTorch are all provided by SageMaker.
- Model training: Develop the object detection model using YOLO, SSD, or Faster R-CNN architectures, using a deep learning framework like TensorFlow or PyTorch. Use the GPUs that SageMaker instances have to speed up training.
- Model evaluation: Assess the trained model using any suitable evaluation metrics, such as mean Average Precision (mAP), Intersection over Union (IoU), or any other.
- Maintaining several versions of the trained models and storing them in S3 or a model registry for quick retrieval and retraining.

3. Deployment Pipeline:
- Model deployment: To deploy the trained model as an API endpoint, use AWS SageMaker's model hosting features. This makes inference scalable and serverless.
- Auto scaling: To optimise cost and resource usage, set up the SageMaker endpoint to automatically scale up or down in response to incoming request volume.
- Monitoring and logging: Use AWS CloudWatch to keep tabs on the deployment model's performance and general health. For the purpose of performance optimisation and troubleshooting, capture and analyse logs.

4. Inference Pipeline:
- Inference infrastructure: To construct a serverless API endpoint for processing inference queries, use API Gateway and AWS Lambda.
- Cost reduction: Use caching mechanisms to cut down on unnecessary inference requests and make use of Amazon CloudFront to cache and serve static assets like photos and models, which will cut down on the overall cost of inference.
- Load balancing: Use Elastic Load Balancing (ELB) to split incoming inference requests amongst many instances, ensuring excellent availability and scalability.

Retraining Approach:

Implement a retraining trigger system based on established parameters like performance decline, the availability of new data, or a predetermined interval for retraining.
- Incremental learning: By initialising the model using the old weights and fine-tuning it on the new data, incremental learning approaches can be used to speed up retraining.
- Automated pipeline: To automatically update the model and release the most recent version, activate the data pipeline, the machine learning pipeline, and the deployment pipeline in order.

Pros, Cons, and Trade-offs:
Utilises managed and scalable AWS services, which eliminates the need for infrastructure administration.
- Supports effective data preprocessing, training, annotation, and storage.
- Facilitates cost reduction through load balancing, resource scalability, and caching.
- Offers a

The top open-source deep learning frameworks and algorithms can be incorporated thanks to the serverless and scalable inference pipeline's flexibility.

Cons and trade-offs: - Because AWS services could be expensive, rigorous cost management and optimisation are necessary.
- Multiple service integration may necessitate more setup and configuration work.
- Additional optimisation and caching techniques might be needed for real-time inference with low latency.

Further Optimization:
Utilise pre-trained models for object detection and fine-tuning with bespoke datasets using AWS AutoML services like Amazon Rekognition.
- Use distributed frameworks like Horovod or numerous instances of distributed training to shorten training periods.
- Use model quantization or compression techniques to optimise inference performance on edge devices and reduce model size.
- Apply serverless batch inference on sizable datasets with AWS Lambda and AWS Batch.

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.

Answer 5 DL:-

1. Data Pipeline:
- Ingest data from current or external sources, such as file systems, databases, or APIs.
- Data storage: Use Azure Blob Storage to safely and permanently store the document data.
- Data preprocessing: Utilise open-source tools like NLTK or SpaCy to perform text preprocessing tasks like tokenization, normalisation, and the elimination of stop words.
- Data labelling: Use pre-existing labelled datasets for supervised training or manually label the documents.
- Document labels or annotations should be stored in Azure Table Storage or Azure Blob Storage.

2. ML Pipeline:

- Training infrastructure: To handle the entire ML workflow, use the Azure Machine Learning service (AML). AML offers tools for model training, deployment, and data preparation.
- Model training: Develop the document categorization model using deep learning models like LSTM or Transformer or machine learning methods like Naive Bayes or Support Vector Machines (SVM).
- Model evaluation: Measures like accuracy, precision, recall, or F1 score can be used to assess the trained model.
- Manage the lifecycle of the trained models and track their various iterations using AML.

3. Deployment Pipeline:
- Model deployment: Use Azure Container Instances (ACI) or Azure Kubernetes Service (AKS) to deploy the trained model. ACI is appropriate for installations of a small scale, but AKS offers scalability and flexibility for bigger workloads.
- Autoscaling: To optimise cost and resource usage, set up AKS to automatically scale up or down in response to incoming inference traffic.
- Performance, health, and resource utilisation of the deployed model are tracked using Azure Monitor and Azure Log Analytics.

4. Inference Pipeline:
- Inference infrastructure: Create serverless functions for handling inference requests using Azure Functions.
- Cost reduction: Use caching techniques to cut down on unnecessary inference requests and make use of the Azure Content Delivery Network (CDN) to cache and provide static content, which will lower the overall cost of inference.
Utilise Azure Traffic Manager or Azure Front Door to load balance inbound inference requests across many instances, guaranteeing high availability and scalability.

Retraining Approach:
Implement a retraining trigger system based on established parameters like performance degradation, the availability of new labelled data, or a predetermined interval for retraining.
- Incremental learning: By initialising the model with past weights and fine-tuning on new labelled data, transfer learning techniques can be used to accelerate retraining.
- Automated pipeline: For smooth updates, trigger the data pipeline, machine learning pipeline, and deployment pipeline sequentially by automating the retraining process with Azure Data Factory or Azure Logic Apps.

Pros, Cons, and Trade-offs:
- Pros:
Utilises managed and scalable Azure services to cut down on infrastructure management costs.
- Offers an integrated end-to-end machine learning workflow using the Azure Machine Learning service.
- Provides flexibility to incorporate different machine learning (ML) methods and methodologies for document classification.
- Facilitates cost optimisation with techniques for load balancing, caching, and autoscaling.

- Cons and trade-offs:
- Because Azure services could be expensive, diligent cost management and optimization are necessary.
- Multiple service integration may necessitate more setup and configuration work.
- Realtime inference using

low-latency may require further optimization and caching strategies.

Further Optimization:
Utilise pre-trained models when classifying and extracting documents with Azure Cognitive Services like Azure Text Analytics or Azure Form Recognizer.
- To speed up training and shorten training time, implement distributed training using Azure Machine Learning service with several nodes or Azure Databricks.
- Use model quantization or compression techniques to optimise inference performance on edge devices and reduce model size.
- Automate the retraining procedure using Azure Logic Apps or Azure Data Factory in response to predetermined triggers.

COMPUTER VISION ASSIGNMENT

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.

Answer 1 CV:-
https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Computer_Vision/CV_1.py

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

Answer 2 CV:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Computer_Vision/CV_2.py**

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

Answer 3 CV:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Computer_Vision/CV_3.py

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

Answer 4 CV:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Computer_Vision/CV_4.py

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

Answer 5 CV:-

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-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 6 CV:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Computer_Vision/CV_6.ipynb

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.

Answer 7 CV:- https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Computer_Vision/CV_7.ipynb

To reduce the inference time of a custom trained segmentation model, you can employ various techniques. Here's an approach you can take:

1. Model Optimization:
    - Quantization: Apply quantization techniques such as post-training quantization or quantization-aware training to reduce the precision of model weights and activations, thereby reducing memory usage and improving inference speed.
    - Pruning: Use pruning techniques to remove unnecessary connections or channels from the model, reducing the model size and inference time.
    - Model architecture optimization: Consider using lighter model architectures that trade off some accuracy for faster inference, such as MobileNet, EfficientNet, or lightweight variants of existing architectures.
2. Inference Optimization:
    - Batch processing: Perform inference on multiple images simultaneously by batching them together. This utilizes parallel processing capabilities of modern hardware and speeds up inference.
    - GPU/CPU optimization: Ensure that your model and data are properly placed on GPUs or CPUs based on their availability and capabilities.
    - Input preprocessing: Optimize image resizing, normalization, and other preprocessing steps to minimize the time spent on these operations during inference.
3. Web App Integration:
    - Create a web app using a web framework like Flask or Django.
    - Implement an image upload functionality in the web app to allow users to upload images.
    - Integrate the custom trained segmentation model into the web app and perform inference on the uploaded images.
    - Display the segmentation predictions to the users and provide the option to download or visualize the results.

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

Answer 8 CV:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Computer_Vision/CV_8.py**

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

Answer 9 CV:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Computer_Vision/CV_9.py**

NATURAL LANGUAGE PROCESSING ASSIGNMENT

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.

Answer 1 NLP:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/NLP_1.py**

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.

Answer 2 NLP:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Natural_Language_Processing/NLP_1.py**

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

Answer 4 NLP:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Natural_Language_Processing/NLP_4.py**

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

Answer 5 NLP:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Natural_Language_Processing/NLP_5.py**

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.

Answer 9 NLP:- **https://github.com/Namitha982000/Placement-Assignment_Namitha-Kolur/blob/master/Natural_Language_Processing/NLP_9.py**