**1 a. With Industry 4.0, artificial intelligence is finding place in every aspect of life. What happens if AI replaces humans in the workplace?**

**ANSWER: (you can write answer which you prepare in class or this below answer.)**

**Positive impact of AI replacing humans in the workplace**:

- Increased efficiency: AI machines can perform repetitive tasks faster and more accurately than humans, leading to increased efficiency and productivity.
- Cost savings: AI machines are more cost-effective in the long run as they do not require salaries, benefits, and time off.
- Improved accuracy: AI machines are programmed to follow a set of instructions, eliminating the possibility of human error.
- 24/7 availability: AI machines can work continuously without rest or breaks, leading to increased productivity and better customer service.

**Negative impact of AI replacing humans in the workplace**:

- Job loss: The widespread use of AI in the workplace could result in massive job losses, particularly in industries where manual labor is prevalent.
- Loss of creativity: AI machines lack the ability to think creatively and make decisions based on intuition.
- Dependence on technology: Over-reliance on AI machines could lead to a lack of critical thinking skills and decision-making abilities in the workforce.
- Ethical concerns: AI machines may be programmed to make decisions that are discriminatory, unethical, or even harmful to humans.

**1 b. for the given scenarios you are required to build an AI solution. Which AI techniques can be applied / best suited for stated problems. Justify**

**1. Extract and digitize the customer information from the Know Your Customer (KYC) forms.**

**2. To identify if employees are wearing face mask in the office campus**

**3. To identify and narrow down tumour regions and further predict if the tumour is malignant or not**

**4. Automated inspection and cost estimation step in the Insurance claim business process**

**5. To identify the location of a moving car within an image**

**ANSWER:**

1. AI techniques best suited for this problem would be Optical Character Recognition(OCR) and Natural Language Processing(NLP). OCR can be used to capture the information from the KYC forms and NLP can be used to digitize the extracted information. OCR can be used to extract the text from the images of the KYC forms and NLP can be used to process the extracted text and convert it into a structured format which can be stored in database. This will help in automating the process of extracting and digitizing customer information from the KYC forms.
2. AI techniques best suited for this problem would be Object Detection using Deep Learning. Convolutional Neural Network(CNNs) can be used to detect objects in images and videos. By training the CNN with labelled images of people wearing masks and not wearing masks, it can be used to accurately identify if an employee is wearing a mask in the office campus.
3. AI techniques best suited for this problem:
   Convolutional Neural Networks(CNNs): CNNs are capable of learning complex features from the images and can be trained to recognize different types of tumours.

Support Vector Machines(SVMs): SVMs can be used to classify the tumour regions as either malignant or benign. SVMs are powerful classifiers can provide good accuracy when trained with sufficient data.

Decision Trees: Decision Trees can be used to determine the risk associated with the tumour. The decision tree can be trained with data on the characteristics of the tumour and can be used to predict the chances of the tumour being malignant.

4. For this scenario, the best AI technique that can be applied is Machine Learning (ML). ML algorithms can be used to develop a model that can learn the patterns in the data and accurately predict the cost estimation for the increase claims. ML can also be used for automated inspection of the claims by analyzing the features such as customer and policy details and post claims to detect any fraudulent claims. This technique can also be used to automatically assign claims to the appropriate teams based on the complexity and urgency of the claim.
5. In this scenario, the best AI technique to be applied is object detection. Object Detection uses computer vision algorithms to detect objects such as cars within an image. This technique can be used to identify the location of a moving car within an image, as it can detect and identify the car in the image and its exact location. This technique can be used in conjunction with other AI techniques such as image segmentation and classification to further improve the accuracy of the detection.

**2 a. Which technique help in addressing certain complex problems with higher accuracy and better generalization characteristics much like human brain in Computer Vision, Natural Language Processing and Speech Domains? And why?**

**ANSWER:** The technique that helps in addressing complex problems in computer vision, natural language processing, and speech domains with higher accuracy and better generalization characteristics is called "Deep Learning".

- Deep Learning algorithms can effectively handle and analyze large amounts of complex data, such as images, videos, text, speech, and complex audio data.
- Deep Learning algorithms are able to model complex relationships in data and extract higher-level features, leading to improved accuracy on various tasks such as image classification, speech recognition, and sentiment analysis.
- Deep Learning algorithms are capable of generalizing to new data, even if it is different from the data they were trained on. This is due to the hierarchical structure of the neural network, which enables the algorithm to learn and extract features at multiple levels.
- With Deep Learning, the process of feature extraction is automated, which saves time and effort compared to manual feature engineering in traditional machine learning.
- Deep Learning algorithms are designed to handle large amounts of data, making them well-suited for tasks such as image recognition and speech recognition, where the amount of data can be huge.
- Deep Learning algorithms are designed to mimic the way the human brain processes information, which makes them ideal for solving problems in domains such as computer vision and natural language processing where the goal is to replicate human-like abilities.

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

- **a. Predicting the food delivery time**
- **b. Predicting whether the transaction is fraudulent**
- **c. Predicting the credit limit of a credit card applicant**
- **d. To group similar customers of an online grocery store, based on their purchasing patterns, to offer discounts to its customers.**
- **e. Predict the probability of a mechanical system breakdown, based on its system vibration and operating temperature**

**ANSWER:**

a. **Predicting the food delivery time**: A regression algorithm such as linear regression or a more complex algorithm such as Random Forest or Gradient Boosting could be applied for this problem. These algorithms can predict a numerical value (delivery time) based on input variables such as location, traffic, and order size.

b. **Predicting whether the transaction is fraudulent**: A classification algorithm such as logistic regression, decision trees, or Random Forest would be best suited for this problem. These algorithms can predict a binary outcome (fraud or non-fraud) based on input variables such as transaction amount, location, and past transaction history.

c. **Predicting the credit limit of a credit card applicant**: A regression algorithm such as linear regression or a more complex algorithm such as Random Forest or Gradient Boosting could be applied for this problem. These algorithms can predict a numerical value (credit limit) based on input variables such as applicant's income, credit history, and employment status.

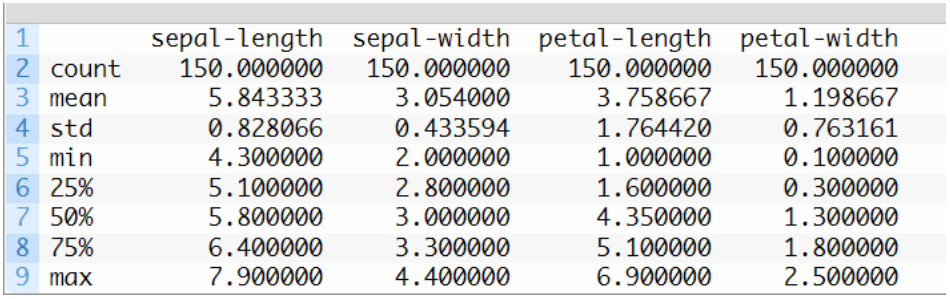
**d. To group similar customers of an online grocery store, based on their purchasing patterns, to offer discounts to its customers**: Clustering algorithms such as K-means or Hierarchical Clustering would be best suited for this problem. These algorithms can group similar customers based on input variables such as purchase history, demographics, and browsing behavior.

**e. Predict the probability of a mechanical system breakdown, based on its system vibration and operating temperature:** A classification algorithm such as logistic regression, decision trees, or Random Forest would be best suited for this problem. These algorithms can predict a binary outcome (breakdown or no breakdown) based on input variables such as vibration and temperature.

**3 a. How to handle the missing values in the dataset? Explain.**

**ANSWER:** [https://www.naukri.com/learning/articles/handling-missing-values-beginners-tutorial/](https://www.naukri.com/learning/articles/handling-missing-values-beginners-tutorial/)

**3 b. The statistical summary of Iris dataset is as follows.**



```
       sepal-length  sepal-width  petal-length  petal-width
count    150.000000   150.000000    150.000000   150.000000
mean       5.843333     3.054000      3.758667     1.198667
std        0.828066     0.433594      1.764420     0.763161
min        4.300000     2.000000      1.000000     0.100000
25%        5.100000     2.800000      1.600000     0.300000
50%        5.800000     3.000000      4.350000     1.300000
75%        6.400000     3.300000      5.100000     1.800000
max        7.900000     4.400000      6.900000     2.500000
```

**Analyse and explain statistical metrics from above summary.**
**ANSWER:**

- Count: count specifies the total number of non-null values in the particular column. For example, the count value for sepal length is 150 specifies total 150 values are there in sepal length column.

- Mean: The mean is a measure of central tendency, indicating the average value of the feature. For example, the mean sepal length is 5.8433 specifies that on average, the sepal length of the iris flowers in this dataset is 5.8433 cm.

- Standard deviation: The standard deviation is a measure of the spread of the data. It indicates how far, on average, the data deviates from the mean. For example, the standard deviation of the sepal length is 0.8280661279778629, meaning that on average, the sepal length of the iris flowers deviates 0.8280661279778629 cm from the mean.

- Minimum: The minimum value is the smallest value in the dataset. For example, the minimum value of sepal length is 4.3, meaning that there is at least one iris flower in the dataset with a sepal length of 4.3 cm.

- Maximum: The maximum value is the largest value in the dataset. For example, the maximum value of sepal length is 7.9, meaning that there is at least one iris flower in the dataset with a sepal length of 7.9 cm.

- Quartiles: The quartiles divide the data into four equal parts, with 25% of the data falling in the first quartile (Q1), 50% of the data falling in the second quartile (Q2), and so on. They give information about the distribution of the data, with the interquartile range (IQR) indicating the spread of the middle 50% of the data.

- 25 Percentile (25%): The 25% specifies the value at which 25% of the value lies below that value. For example, the 25% value for sepal length is 5.1 which indicate that 25% of the total values of sepal length are below 5.1.

- 50 Percentile (50%): The 50% specifies the value at which 50% of the value lies below that value. For example, the 50% value for sepal length is 5.8 which indicate that 50% of the total values of sepal length are below 5.8.

- 75 Percentile (75%): The 75% specifies the value at which 75% of the value lies below that value. For example, the 75% value for sepal length is 6.4 which indicate that 75% of the total values of sepal length are below 6.4.

**4 a. Consider a real estate company that has a dataset containing the prices of properties in the Delhi region. It wishes to use the data to optimise the sale prices of the properties based on important factors such as area, bedrooms, parking, etc.**

**Essentially, the company wants —**

a. **To identify the variables affecting house prices, e.g. area, number of rooms, bathrooms, etc.**
b. **To create a model that quantitatively relates house prices with variables such as number of rooms, area, number of bathrooms, etc.**
c. **To know the accuracy of the model, i.e. how well these variables can predict house prices.**

**Discuss the steps to be followed to build such a model. Recommend the suitable techniques to consider at each step.**

**ANSWER:**

**Step 1: Data collection or load the dataset**:

- Collect and gather all relevant data, including house prices and the variables affecting house prices such as area, number of rooms, bathrooms.
- To load the dataset. The data can be loaded from different website like **kaggle, sklearn**, etc.

**Step 2: Data cleaning and preprocessing**:

- Clean and preprocess the data to handle missing values, outliers, and ensure data is in a suitable format for analysis.
- To check for missing values we use pandas **isnull().sum()** function, and to remove the missing values we can use **impute() or interpolate()** function.
- To handle the outliers we can use **Box Plot or IQR method**.

**Step 3: Data Transformation:**

- Data transformation is used to convert data from one format to another.
- Dataset may contain the categorical columns which needs to be converted into numerical value. To do this we can use the **OneHotEncoder(), or OrdinalEncoder().**
- Here in this example we can user **MinMaxScaler(), StandardScaler().**

**Step 4: Exploratory Data Analysis (EDA):**

- Perform exploratory data analysis to understand the relationship between variables and identify any trends, patterns, or correlations.
- Here in example we can use the **pair plot or Heatamap** to fine the relationship between input variables number of rooms, area, number of bathrooms and output variable house prices**.**

**Step 5: Feature selection:**

- Choose the most relevant variables to be used in the model by analyzing the correlation between variables and the target (property prices).
- Here in example, from set of input variable we can select the input features are not correlated or which plays import role in calculating House price.
- To select the best features we can also use the **Principal Component Analysis (PCA**)

**Step 6: Model building:**

- Choose a suitable algorithm such to build the model and estimate the relationship between variables and property prices.

- For the given problem the regression algorithm such as **multiple linear regression**, can be used to build the model and estimate the relationship between input variables and House prices.

**Step 7:Model evaluation:**

- Evaluate the model's performance using metrics such as **mean squared error, R-squared, and cross-validation.**

**Step 8: Model tuning:**

- Fine-tune the model to improve its accuracy by adjusting its parameters and hyperparameters.

**Step 9: Model deployment:**

- Deploy the final model on local server or cloud server for practical use and make predictions on House Price.

**Reference: https://www.kaggle.com/code/ashydv/housing-price-prediction-linear-regression**

**4b. Describe univariate, bivariate, and multivariate analysis with suitable examples.**
**ANSWER:**
**Univariate analysis**:

- Univariate analysis is the simplest form of data analysis that deals with only one variable at a time.
- It involves summarizing and describing the properties of a single variable, such as its central tendency (mean, median, mode), spread (standard deviation, range), and shape (skewness, kurtosis).
- For example, if we want to know the average price of houses in a city, we can use univariate analysis to analyze the "price" variable.
- To perform a univariate analysis, one could calculate the mean, median, and mode of the heights, and create a histogram to visualize the distribution.
- The histogram, count plot, and box plot are commonly used plots in univariate analysis as it helps to understand the shape of the distribution, count number of value, and identify any outliers.

**Bivariate analysis**:

- Bivariate analysis, on the other hand, involves analyzing the relationship between two variables.
- It helps to understand how changes in one variable affect the other variable.
- For example, if we want to know the relationship between the size of a house and its price, we can use bivariate analysis to analyze the relationship between the "size" and "price" variables.
- To perform a bivariate analysis, one could calculate the correlation coefficient and covariance between the two variables and create a scatter plot to visualize the relationship.
- The scatter plot and heat map are commonly used plots in bivariate analysis as it helps to identify any patterns or trends in the data, find correlation, and determine the strength of the relationship between the variables.

**Multivariate analysis**:

- Multivariate analysis involves analyzing more than two variables at a time.
- It helps to understand the relationships between multiple variables and identify any patterns or trends in the data.

- For example, if we want to know the relationship between the size, number of bedrooms, and price of a house, we can use multivariate analysis to analyze the relationship between the "size", "number of bedrooms" and "price" variables.
- To perform a multivariate analysis, one could calculate the correlation matrix between all variables and create a three-dimensional scatterplot to visualize the relationship between all three variables.
- The three-dimensional scatterplot is a commonly used plot in multivariate analysis as it helps to visualize the relationships between multiple variables in a single plot. However, other plots such as heatmaps and parallel coordinates can also be used to perform a multivariate analysis.

**5 a. N-grams are defined as the combination of N keywords together. Consider the given sentence: "Data Visualization is a way to express your data in a visual context so that patterns, correlations, trends between the data can be easily understood."**
**Generate bi-grams and tri-grams for the above sentence**

**a. Before performing text cleaning steps.**

**b. After performing following text cleaning steps:**

**1. Stop word Removal**

**2. Replacing punctuations by a single space**

**ANSWER:**

**a. Before performing text cleaning steps.**

**Bigram**

[(Data, Visualization), (Visualization, is), (is, a), (a, way), (way, to), (to, express), (express, your), (your, data), (data, in), (in, a), (a, visual), (visual, context), (context, so), (so, that), (that, patterns), (patterns, ,), (,, correlations), (correlations, ,), (,, trends), (trends, between), (between, the), (the, data), (data, can), (can, be), (be, easily), (easily, understood)]

**Trigram:**

[(Data, Visualization, is), (Visualization, is, a), (is, a, way), (a, way, to), (way, to, express), (to, express, your), (express, your, data), (your, data, in), (data, in, a), (in, a, visual), (a, visual, context), (visual, context, so), (context, so, that), (so, that, patterns), (that, patterns, ,), (patterns, ,, correlations), (,, correlations, ,), (correlations, ,, trends), (,, trends, between), (trends, between, the),(between, the, data), (the, data, can), (data, can, be), (can, be, easily), (be, easily, understood)]

**b. After performing following text cleaning steps:**

**1. Stop word Removal**

**2. Replacing punctuations by a single space**

**Bigram**

[(Data, Visualization), (Visualization, way), (way, express), (express, data), (data, visual), (visual, context), (context, patterns), (patterns, correlations), (correlations, trends), (trends, data), (data, easily), (easily, understood)]

**Trigram:**

[(Data, Visualization, way), (Visualization, way, express), (way, express, data), (express, data, visual), (data, visual, context), (visual, context, patterns), (context, patterns, correlations), (patterns, correlations, trends), (correlations, trends, data), (trends, data, easily), (data, easily, understood)]

**5b. K-means clustering with Euclidean distance suffer from the curse of dimensionality. Is the statement true and why?**

**ANSWER:**

The statement is true. The curse of dimensionality refers to the fact that as the number of dimensions in a dataset increases, the amount of data required to accurately model the data also increases exponentially.

- K-mean clustering with Euclidean distance is sensitive to outliers, which is amplified when the dimensionality of the data increases. With high number of dimensions, the distance between points in the dataset can become much larger than the distance between points in the same cluster. This makes it difficult for the algorithm to accurately identify cluster and can lead to suboptimal clustering result.
- Additionally, when the dimensionality increases, the number of data points needed for accurate clustering also increases exponentially, making it difficult to achieve accurate result.
- High dimensional datasets also tend to be sparse, resulting in clustering that are not tightly distributed.
- Finally, the effectiveness of Euclidean distance as a clustering metric decreases with higher dimensionality because the Euclidean distance between points in higher dimensions becomes less meaningful.

**6a. The sinking of the Titanic is one of the most infamous shipwrecks in history.**
**On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. You are asked to build a machine learning model to predict whether a passenger survived or not. Describe each step you will follow to build this model.**

**ANSWER: (you can write the answer which written in CIE or this below answer)**

**1.Data Exploration:** The first step is to explore the dataset and understand the features, their distribution, and any missing values. This can be done by using various data visualization techniques such as histograms, box plots, and scatter plots.

**2. Data Preprocessing**: After exploring the dataset, the next step is to preprocess the data, which includes cleaning and transforming the data to make it ready for modeling. This step involves handling missing values, encoding categorical variables, and normalizing the data.

**3. Feature Selection**: In this step, we select the most relevant features for the model. This can be done by using various techniques such as correlation analysis, mutual information, and chi-squared test.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# load data
data = pd.read_csv("titanic.csv")
```

```
# handle missing values
data = data.fillna(data.mean())

# convert categorical variables to numerical
data = pd.get_dummies(data, columns=["Sex", "Embarked"])

# plot heatmap to visualize correlations
sns.heatmap(data.corr(), annot=True)
plt.show()

# select relevant features
features = ["Pclass", "Age", "Fare", "Sex_male", "Embarked_C", "Embarked_Q"]
target = "Survived"
X = data[features]
y = data[target]
```

**4. Model Selection** and splitting train and test dataset: After selecting the features, split the dataset set into train and test set. Then select the appropriate machine learning model. In this case, since it's a classification problem, we can use various classification algorithms like logistic regression, decision tree, Random Forest, Naive Bayes, etc.

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_state=42)
from sklearn.tree import DecisionTreeClassifier
dtc=DecisionTreeClassifier()
```

**5. Model Training**: Once the model is selected, it is trained using the selected features and the target variable.

```
dtc.fit(x_train,y_train)
pred=dtc.predict(x_test)
pred
```

**6. Model Evaluation**: After training the model, it is evaluated using various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

```
from sklearn.metrics import accuracy_score,confusion_matrix
print('accuracy of decision tree',accuracy_score(y_test,pred))
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)
```

**7. Model Tuning**: If the model's performance is not satisfactory, we can tune the model's hyperparameters to improve its performance.

**8. Model Deployment**: After the model is trained and tuned, it is deployed in a production environment to make predictions on new, unseen data.

**9. Model Maintenance**: Finally, model performance should be monitored regularly and updated as needed.

**6b. You work for a textile manufacturer and have been asked to build a model to detect and classify fabric defects. You trained a machine learning model with high recall. You want quality control inspectors to gain trust in your model. Which technique should you use to understand the rationale of your classifier? Justify**

**ANSWER**: Integrated Gradients method to efficiently compute feature attributions for each predicted image.

- Integrated Gradients is a machine learning interpretation method that helps to detect and classify fabric defects by attributing a portion of the prediction of a deep learning model to each pixel or feature in an image. The method provides a visual representation of the importance of each pixel or feature in the prediction and helps to identify the specific regions of an image that are contributing to a classification.

- For detecting and classifying fabric defects, the Integrated Gradients method can be applied to the input image of a deep learning model that has been trained on fabric images with and without defects. The method will highlight the regions of the image that are contributing most to the prediction of the presence of a defect.

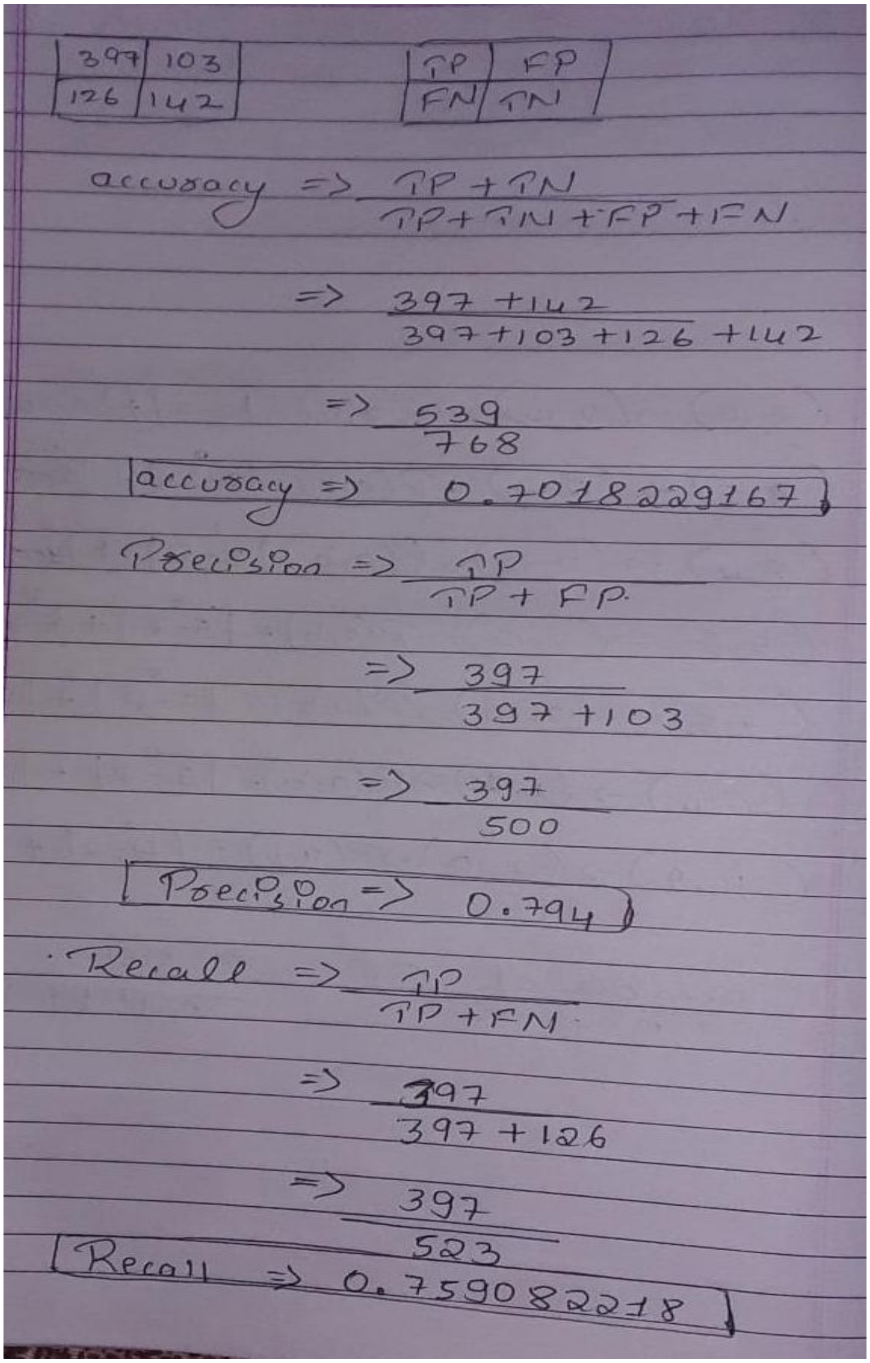
- This information can be used to identify the specific types of defects, such as holes, tears, or stains, by comparing the regions of the image with a library of known defect patterns. The Integrated Gradients method can also be used to determine the extent to which different types of defects are affecting the final prediction, allowing for more accurate and precise classification.

- Overall, the Integrated Gradients method can provide a valuable tool for the detection and classification of fabric defects, allowing for the quick and efficient identification of problems in the manufacturing process.

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

|  |  | Actual |  |
|---|---|---|---|
|  |  | 1 | 0 |
| Predicted | 1 | 397 | 103 |
|  | 0 | 126 | 142 |

**ANSWER:**



|  |  |
|---|---|
| 397 | 103 |
| 126 | 142 |

|  |  |
|---|---|
| TP | FP |
| FN | TN |

accuracy $\Rightarrow \dfrac{TP+TN}{TP+TN+FP+FN}$

$\Rightarrow \dfrac{397+142}{397+103+126+142}$

$\Rightarrow \dfrac{539}{768}$

$\boxed{\text{accuracy} \Rightarrow 0.7018229167}$

Precision $\Rightarrow \dfrac{TP}{TP+FP}$

$\Rightarrow \dfrac{397}{397+103}$

$\Rightarrow \dfrac{397}{500}$

$\boxed{\text{Precision} \Rightarrow 0.794}$

Recall $\Rightarrow \dfrac{TP}{TP+FN}$

$\Rightarrow \dfrac{397}{397+126}$

$\Rightarrow \dfrac{397}{523}$

$\boxed{\text{Recall} \Rightarrow 0.759082218}$

Specificity :- $\frac{TN}{TN + FP}$

$$\Rightarrow \frac{142}{142 + 103}$$

$$\Rightarrow \frac{142}{245}$$

$\boxed{\text{Specificity} \Rightarrow 0.5795918367}$

F1-Score $\Rightarrow 2\left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$

$$2.\left(\frac{0.794 \ * \ 0.759}{0.794 \ + \ 0.759}\right)$$

$$2\left(\frac{0.602646}{1.553}\right)$$

$$2.\left(0.388052801\right)$$

$\boxed{\text{F1-Score} \Rightarrow 0.7761056021}$

AUC-ROC $\Rightarrow$

**7b. A Machine Learning Engineer is preparing a data frame for a supervised learning task. The ML Engineer notices the target label classes are highly imbalanced and multiple feature columns contain missing values. The proportion of missing values across the entire data frame is less than 5%. What should the ML Engineer do to minimize bias due to missing values? Support your argument.**

**ANSWER:**

- Remove rows with missing values: The simplest approach, but can result in loss of information if the proportion of missing values is high.
- Impute missing values: Using statistical methods such as mean, median, mode or regression to estimate missing values. However, it may not accurately reflect the actual value and introduce bias.
- Use advanced imputation methods: Such as Multiple Imputation, which creates multiple imputed datasets and combines the results for a more accurate estimate.
- Use a separate supervised learning model for imputing missing values: Building a model specifically for imputing missing values can result in a more accurate estimate.

Therefore,

Use supervised learning to predict missing values based on the values of other features. Different supervised learning approaches might have different performances, but any properly implemented supervised learning approach should provide the same or better approximation than mean or median approximation. (**very important point**)

**8 a. A data scientist is working on optimizing a model during the training process by varying multiple parameters. The data scientist observes that, during multiple runs with the identical parameters the loss function converges to different, yet stable values. What should the data scientist do to improve the training process? Justify.**

**ANSWER**: As a data scientist we can ***reduce the batch size and decrease the learning rate***. Because of the following reasons:

- Smaller batch sizes allow the model to learn more diverse information from the data, leading to better generalization on unseen data.
- When the batch size is small, the model's parameters are updated more frequently, leading to a more fine-tuned optimization process.
- A lower learning rate enables the model to converge more slowly and precisely to the optimal solution.
- By learning more slowly, the model is less likely to get stuck in poor local minima and can find a better global minimum.
- A lower learning rate reduces the size of the steps taken by the optimization algorithm, reducing the chances of overshooting the optimal solution.

It is most likely that the loss function is very curvy and has multiple local minima where the training is getting stuck. Decreasing the batch size would help the data scientist stochastically get out of the local minima saddles. Decreasing the learning rate would prevent overshooting the global loss function minimum.

**8 b. A company has collected customer comments on its products, rating them as safe or unsafe, using decision trees. The training dataset has the following features: id, date, full review, full review summary, and a binary safe/unsafe tag. During training, any data sample with missing features was dropped. In a few instances, the test set was found to be missing the full review text field. For this use case, which is the most effective course of action to address test data samples with missing features. Justify**

**ANSWER:** Missing values are present in the "**full review text field**", so we have to impute missing values using Natural Language Processing (NLP), because the "**full review text field"** column contains the data in the form of text.

Imputation using NLP techniques is a popular method to handle missing values in the text column of a dataset because:

- **Preserves information content**: By using NLP techniques to estimate missing text, the original information content of the text column can be preserved to a large extent, ensuring better representation of the data.
- **Better accuracy**: NLP techniques are designed to work with text data, which means that they can provide more accurate estimates of missing text compared to other imputation methods.
- **Avoids loss of information**: By imputing missing values, important information is not lost, which would be the case if missing values were simply dropped.
- **Better generalization performance**: By preserving the information content of the text column, the model can generalize better to new data.

**9a. what are the deployment strategies borrowed from DevOps that can be utilized in MLOPs. Explain anyone strategy.**

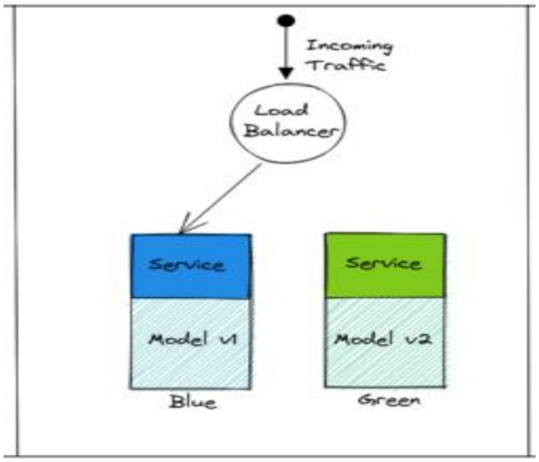
**ANSWER:** There are several deployment strategies borrowed from DevOps that can be utilized in MLOps (Machine Learning Operations). Some of the common strategies are:

1. **Continuous Integration and Continuous Deployment (CI/CD**): The practice of automating the build, test, and deployment of machine learning models using a pipeline.
2. **Blue-Green Deployment**: A deployment strategy where two identical production environments are maintained, with only one being active at a time. The inactive environment can be updated with the new version and tested, and then swapped with the active environment.
3. **Canary Deployment**: A gradual rollout of the new version to a small subset of users or systems, before rolling it out to the entire infrastructure.
4. **A/B Testing**: A technique for comparing two or more variations of a machine learning model to determine which one performs better.

**Blue-Green Deployments** :

- Blue-Green Deployments is a deployment strategy that aims to minimize downtime and reduce the risk of deployment errors. This strategy involves having two parallel production environments, one known as the "blue" environment and the other known as the "green" environment.
- During normal operation, the "blue" environment is live and serving the production traffic. When a new version of the application or a new model is ready to be deployed, it is deployed to

the "green" environment. The "green" environment is then thoroughly tested to ensure that it is working as expected.



- Once the "green" environment has been validated, the traffic is redirected from the "blue" environment to the "green" environment, making the new version live. The "blue" environment is then updated and becomes the new backup environment.
- This process can be repeated for subsequent deployments, with the "green" environment becoming the new "blue" environment and the "blue" environment being updated.
- The benefits of using the blue-green deployment strategy are that it minimizes downtime, as the switchover between environments is quick and seamless. It also reduces the risk of deployment errors, as the new version can be thoroughly tested before it goes live. Additionally, it provides an easy rollback option, as the previous version is still available in the "blue" environment.

**9 b. Machine learning models can be resource heavy. They require a good amount of processing power to predict, validate, and recalibrate, millions of times over. How can containerization of ML model solve this problem?**

**ANSWER:**
Containerization is a method of packaging software into containers that include everything required to run the software, including libraries, dependencies, and system tools. This process helps to isolate software from the underlying infrastructure, making it easier to deploy and run applications in a consistent and predictable manner.

Containerization helps to solve the problem of resource heavy machine learning models in several ways:

- Portability: Containers provide a consistent environment that can be easily moved from one system to another, without any compatibility issues. This makes it easier to run the models on different platforms and infrastructure, reducing the need for specialized hardware or expensive cloud services.
- Scalability: Containers allow for the creation of multiple instances of the same model, which can run in parallel. This can help to scale the model to handle larger volumes of data and more complex computations, without incurring additional hardware costs.

- Resource Management: Containers provide an isolated environment that can be configured to run efficiently within a specific resource constraint, such as memory, CPU, or network bandwidth. This can help to optimize the performance of the model, and reduce the amount of resources needed to run it.
- Reproducibility: Containerization helps to ensure that the model runs in a consistent and predictable environment, which can help to minimize the risk of errors or inconsistencies in the model's predictions. This can improve the accuracy and reliability of the model over time, as it is recalibrated and validated.
- Security: Containers provide a secure environment that can be isolated from the rest of the system, reducing the risk of security breaches or data leaks. This can help to protect sensitive information and data, and ensure that the model is used only for its intended purpose.

**10 a. how will you deploy a trained machine learning model as a predictive service in a production environment. Explain.**

**ANSWER:**

- First, the trained machine learning model will need to be exported or serialized in a format that can be easily loaded and used in a production environment. This may include formats such as pickle, joblib, or TensorFlow SavedModel.
- Next, the model will need to be deployed on a production server or cloud platform that can handle the processing and prediction requests. This could include platforms such as Amazon SageMaker, Google Cloud ML Engine, or Microsoft Azure Machine Learning.
- Once the model is deployed, it will need to be integrated into a web or mobile application that can make predictions based on user input. This could include an API endpoint that takes in input data and returns the predicted output.
- To ensure the model is running smoothly and accurately, monitoring and logging tools will need to be implemented. This will allow for tracking of usage, prediction accuracy, and any errors that may occur.
- Finally, the model will need to be updated and retrained as necessary to ensure it remains accurate and performs well in the production environment. This may involve incorporating feedback from users, monitoring performance metrics, and incorporating new data.

**10 b. for the below given scenarios, suggest best suited cloud deployment model and list the challenges with it.**

**1. for, a. Variable workload b. Test and Development**

**2. for, a. Cloud bursting, b. On demand access, c. Sensitive data**

**ANSWER:**

1. The best suited cloud deployment model is hybrid cloud.

**Challenges**:

- Security and compliance issue due to data being stored in multiple location.
- Difficulty in managing dispersed cloud environment.
- Complexity of integration between services from different cloud provider.

2. The best suited cloud deployment model is private cloud.

**Challenges**:

- Higher initial cost for setting up and maintaining the private cloud infrastructure
- Limited scalability compared to public cloud options
- Potential security risks if not properly configured and managed.

**OR**

**ANSWER**:

1.a. Best suited cloud deployment model is Public Cloud.

**Challenges**:

- Security concerns with sharing resources with other organizations on the same public cloud platform
- Limited control over infrastructure and potential disruptions due to maintenance or updates by the cloud provider
- Potential for higher costs due to increased usage during peak periods or increased data storage needs
- Difficulty in migrating existing applications and data to the public cloud
- Potential compliance issues with regulations and industry standards.

1.b. Best suited cloud deployment model is Private Cloud

**Challenges:**

- Higher initial cost for setting up and maintaining the private cloud infrastructure
- Limited scalability compared to public cloud options
- Dependence on the organization's internal IT resources for management and maintenance
- Potential security risks if not properly configured and managed.

2. a. Cloud Deployment Model is Hybrid Cloud.

**Challenges**:

- Integration and compatibility issues between different cloud environments
- Complexity in managing and maintaining multiple cloud environments
- Difficulty in ensuring security and compliance across different cloud environments
- Potential for increased costs due to the need to manage multiple cloud environments
- Difficulty in ensuring consistent performance and availability across different cloud environments.

2.b.

**Scenario 1**: A small business wants to host their website and store customer data

Best suited cloud deployment model is Public Cloud

**Challenges**:

- Security concerns around storing sensitive customer data on a public cloud
- Limited control over the infrastructure and potential for service outages
- Potential for additional costs if usage exceeds the initial budget

**Scenario 2**: A large enterprise wants to host their ERP system for employees to access remotely
Best suited cloud deployment model is Private Cloud
**Challenges**:

- Higher initial costs for setting up and maintaining a private cloud infrastructure
- Limited flexibility and scalability compared to a public cloud
- Difficulty in managing and maintaining a secure and reliable infrastructure
- Dependence on a single vendor for support and maintenance

**Scenario 3**: A startup wants to launch a new app quickly and scale it as needed
Best suited cloud deployment model is Hybrid Cloud
**Challenges**:

- Complexity in managing and integrating multiple cloud environments
- Difficulty in ensuring data security and compliance across different cloud environments
- Potential for increased costs if the startup does not effectively manage and optimize its usage of the hybrid cloud environment.

2.c. Cloud deployment model is Private Cloud
**Challenges**:

- Ensuring the security and privacy of sensitive data.
- Maintaining compliance with regulatory requirements for handling sensitive data.
- Managing access controls and permissions for sensitive data.
- Ensuring data integrity and availability in the event of a disaster or system failure.
- Managing the costs associated with maintaining a private cloud infrastructure.