Q1: What are missing values in a dataset? Why is it essential to handle missing values? Name some algorithms that are not affected by missing values.

Missing Data that's Missing Completely at Random (MCAR) : These are data that are missing completely at random. That is, the missingness is independent from the data. There is no discernible pattern to this type of data missingness.

Missing Data that's Missing at Random (MAR) : These types of data are missing at random but not completely missing. The data's missingness is determined by the data you see.

Missing Data that's Not Missing at Random (NMAR): These are data that are not missing at random and are also known as ignorable data. In other words, the missingness of the missing data is determined by the variable of interest.

Why is it essential to handle missing values : Missing data are problematic because, depending on the type, they can sometimes bias your results. This means your results may not be generalisable outside of your study because your data come from an unrepresentative sample.

Q2: List down techniques used to handle missing data. Give an example of each with python code.

1.Delete Rows with Missing Values

2.Impute missing values with Mean/Median:

3.Imputation method for categorical columns:

In [3]:

```
import seaborn as sns
df = sns.load_dataset('titanic')
df.head()
```

Out[3]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_ma
0	0	3	male	22.0	1	0	7.2500	S	Third	man	Tru
1	1	1	female	38.0	1	0	71.2833	C	First	woman	Fals
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	Fals
3	1	1	female	35.0	1	0	53.1000	S	First	woman	Fals
4	0	3	male	35.0	0	0	8.0500	S	Third	man	Tru

In [4]:

```
# delete rows
df.dropna()

# impute missing value mean/median
df['age'].fillna(df['age'].mean())

# impute missing categorial
mode = df[df['embarked'].notna()]['embarked'].mode()[0]
df['embarked'].fillna(mode)
```

Out[4]:

```
0      S
1      C
2      S
3      S
4      S
      ..
886    S
887    S
888    S
889    C
890    Q
Name: embarked, Length: 891, dtype: object
```

Q3: Explain the imbalanced data. What will happen if imbalanced data is not handled?

An imbalanced dataset means instances of one of the two classes is higher than the other, in another way, the number of observations is not the same for all the classes in a classification dataset.

This problem is faced not only in the binary class data but also in the multi-class data.

The problem with training the model with an imbalanced dataset is that the model will be biased towards the majority class only. This causes a problem when we are interested in the prediction of the minority class \

if dataset is imbalance then model will bias toward the one result

Q4: What are Up-sampling and Down-sampling? Explain with an example when up-sampling and down-sampling are required.

Downsampling (in this context) means training on a disproportionately low subset of the majority class examples.

Upweighting means adding an example weight to the downsampled class equal to the factor by which you downsampled.

If there are two classes, then balanced data would mean 50% points for each of the class. For most machine learning techniques, little imbalance is not a problem. So, if there are 60% points for one class and 40% for the other class, it should not cause any significant performance degradation. Only when the class imbalance is high, e.g. 90% points for one class and 10% for the other, standard optimization criteria or performance measures may not be as effective and would need modification.

A typical example of imbalanced data is encountered in e-mail classification problem where emails are classified into ham or spam. The number of spam emails is usually lower than the number of relevant (ham) emails. So, using the original distribution of two classes leads to imbalanced dataset.

Q5: What is data Augmentation? Explain SMOTE.

Data augmentation is a set of techniques to artificially increase the amount of data by generating new data points from existing data. This includes making small changes to data or using deep learning models to generate new data points.

SMOTE : Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in your dataset in a balanced way. The component works by generating new instances from existing minority cases that you supply as input. This implementation of SMOTE does not change the number of majority cases.

The new instances are not just copies of existing minority cases. Instead, the algorithm takes samples of the feature space for each target class and its nearest neighbors. The algorithm then generates new examples that combine features of the target case with features of its neighbors. This approach increases the features available to each class and makes the samples more general.

SMOTE takes the entire dataset as an input, but it increases the percentage of only the minority cases. For example, suppose you have an imbalanced dataset where just 1 percent of the cases have the target value A (the minority class), and 99 percent of the cases have the value B. To increase the percentage of minority cases to twice the previous percentage, you would enter 200 for SMOTE percentage in the component's properties.

Q6: What are outliers in a dataset? Why is it essential to handle outliers?

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations

Outliers are important because they can have a large influence on statistics derived from the dataset. For example, the mean intake of energy or some nutrient may be [glossary term:]skewed upward or downward by one or a few extreme values

Q7: You are working on a project that requires analyzing customer data. However, you notice that some of the data is missing. What are some techniques you can use to handle the missing data in your analysis?

If there Most of of value is Missing in any partucular column then i will delete that column. if column is important for model than i can impute mean of column if there is no outlier in column otherwise i may use median. if mising value is present in categorial column then i will put mode value

Q8: You are working with a large dataset and find that a small percentage of the data is missing. What are some strategies you can use to determine if the missing data is missing at random or if there is a pattern to the missing data?

i may impute missing data with mean or median

Q9: Suppose you are working on a medical diagnosis project and find that the majority of patients in the dataset do not have the condition of interest, while a small percentage do. What are some strategies you can use to evaluate the performance of your machine learning model on this imbalanced dataset?

i will upsample theminority class to provide more data of miority class to model so that model can work better

Q10: When attempting to estimate customer satisfaction for a project, you discover that the dataset is unbalanced, with the bulk of customers reporting being satisfied. What methods can you employ to balance the dataset and down-sample the majority class?

i shouldn't down sample the majority class becuase if i down sample the class it most likely than important data will be lost.so think i should upsample the customer unsatisfied reviews.

Q11: You discover that the dataset is unbalanced with a low percentage of occurrences while working on a project that requires you to estimate the occurrence of a rare event. What methods can you employ to balance the dataset and up-sample the minority class?

In [ ]: