PYTHON

**How would you handle an imbalanced dataset?**

**Answer:**An imbalanced dataset is when you have, for example, a classification test and 90% of the data is in one class. That leads to problems: an accuracy of 90% can be skewed if you have no predictive power on the other category of data! Here are a few tactics to get over the hump:

1. Collect more data to even the imbalances in the dataset
2. Resample the dataset to correct for imbalances.
3. Oversampling
4. Undersampling1
5. Try a different algorithm altogether on your dataset.

What’s important here is that you have a keen sense for what damage an unbalanced dataset can cause, and how to balance that.

#### How do you handle missing or corrupted data in a dataset?

**Answer:**You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value (0, -9999, mean, median)

In Pandas, there are two very useful methods: isnull() and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

#### Which data visualization libraries do you know in python? Mention the popular two

**Answer:**seaborn and matplotlib

**What is Machine learning?**

Machine learning is a branch of computer science which deals with system programming in order to automatically learn and improve with experience.  For example: Robots are programed so that they can perform the task based on data they gather from sensors. It automatically learns programs from data.

**Mention 5 popular algorithms for supervised learning problem**

* Decision Trees
* Neural Networks (back propagation)
* Probabilistic networks
* K- Nearest Neighbor
* Support vector machines
* Decision Tree and Random Forest

**What are the different problems in Machine Learning?**

The different types of techniques in Machine Learning are

* Supervised Learning
* Unsupervised Learning
* Semi-supervised Learning
* Reinforcement Learning

**What is ‘Training set’ and ‘Test set’?**

In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as ‘Training Set’. Training set is an examples given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner. Training set are distinct from Test set.

**Write a code to implement K-Nearest Neighbor and instantiate it?**

#using knn

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier()

**Give a popular application of machine learning that you see on day-to-day basis?**

The recommendation engine implemented by major ecommerce websites uses Machine Learning is an example.

Explain Free Launch Theorem

Says no way to know a model that will perform better until you test it.