11. How do you handle missing data?

Answer :

Handling Missing Data:

- Removal: Delete rows or columns with missing values.
- Imputation: Fill in missing values with mean, median, mode, or using algorithms like KNN.
- Prediction: Use models to predict and replace

missing values.

12. What are some common algorithms for clustering?

Answer :

Common Clustering Algorithms:

- K-Means: Partitions data into K clusters based on the mean distance.
- Hierarchical Clustering: Builds a hierarchy of

clusters using a tree-like structure.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Finds clusters based on the density of data points.

13. What is the difference between correlation and causation?

Answer :

Correlation:

- Measures the strength and direction of the relationship between two variables.
- Correlation does not imply causation.

Causation:

- Indicates that one event is the result of the occurrence of the other event; there is a cause-and-effect relationship.

14. Explain the Central Limit Theorem (CLT) and its significance.

Answer :

Central Limit Theorem (CLT):

- States that the distribution of the sample mean of a sufficiently large number of independent and identically distributed

(i.i.d.) variables will approximate a normal distribution, regardless of the original distribution of the population.
- Significance: Allows for the use of normal distribution properties in inferential statistics, such as confidence intervals and hypothesis testing.

15. What are some techniques for handling

imbalanced datasets?

Answer :

Techniques for Handling Imbalanced Datasets:

- Resampling: Over-sampling the minority class or under-sampling the majority class.
- Synthetic Data Generation: Using techniques like SMOTE (Synthetic Minority

Over-sampling Technique).

- Anomaly Detection: Treating the minority class as anomalies.
- Ensemble Methods: Using algorithms like Random Forest or boosting that can handle imbalance.
- Adjusting Class Weights: Assigning higher weights to the minority class during training.

16. Explain the concept of feature engineering and its importance.

Answer :

Feature Engineering:

- The process of creating new features or modifying existing features to improve the performance of machine learning models.
- Importance: Helps in

providing better inputs to the model, thus improving accuracy and predictive power.

17. What is the purpose of regularization in machine learning?

Answer :

Regularization:

- A technique used to prevent overfitting by adding a penalty term to the model's loss function.

- Types:
  - L1 Regularization (Lasso): Adds the absolute value of coefficients as penalty.
  - L2 Regularization (Ridge): Adds the squared value of coefficients as penalty.

18. How do you choose the number of clusters in K-means clustering?

Answer :

Choosing the Number of Clusters:

- Elbow Method: Plot the within-cluster sum of squares (WCSS) against the number of clusters and look for the "elbow" point.
- Silhouette Score: Measures how similar an object is to its own cluster compared to other

clusters.

- Gap Statistic: Compares the total within intra-cluster variation for different numbers of clusters with their expected values under null reference distribution of the data.

19. Explain the difference between PCA and LDA.

Answer :

Principal Component Analysis (PCA):

- A dimensionality reduction technique that projects data onto the directions of maximum variance.
- Unsupervised learning method.

Linear Discriminant Analysis (LDA):

- A classification and dimensionality reduction technique that projects data to maximize the separation between classes.

- Supervised learning method.

20. What is the difference between a ROC curve and a Precision-Recall curve?

Answer :

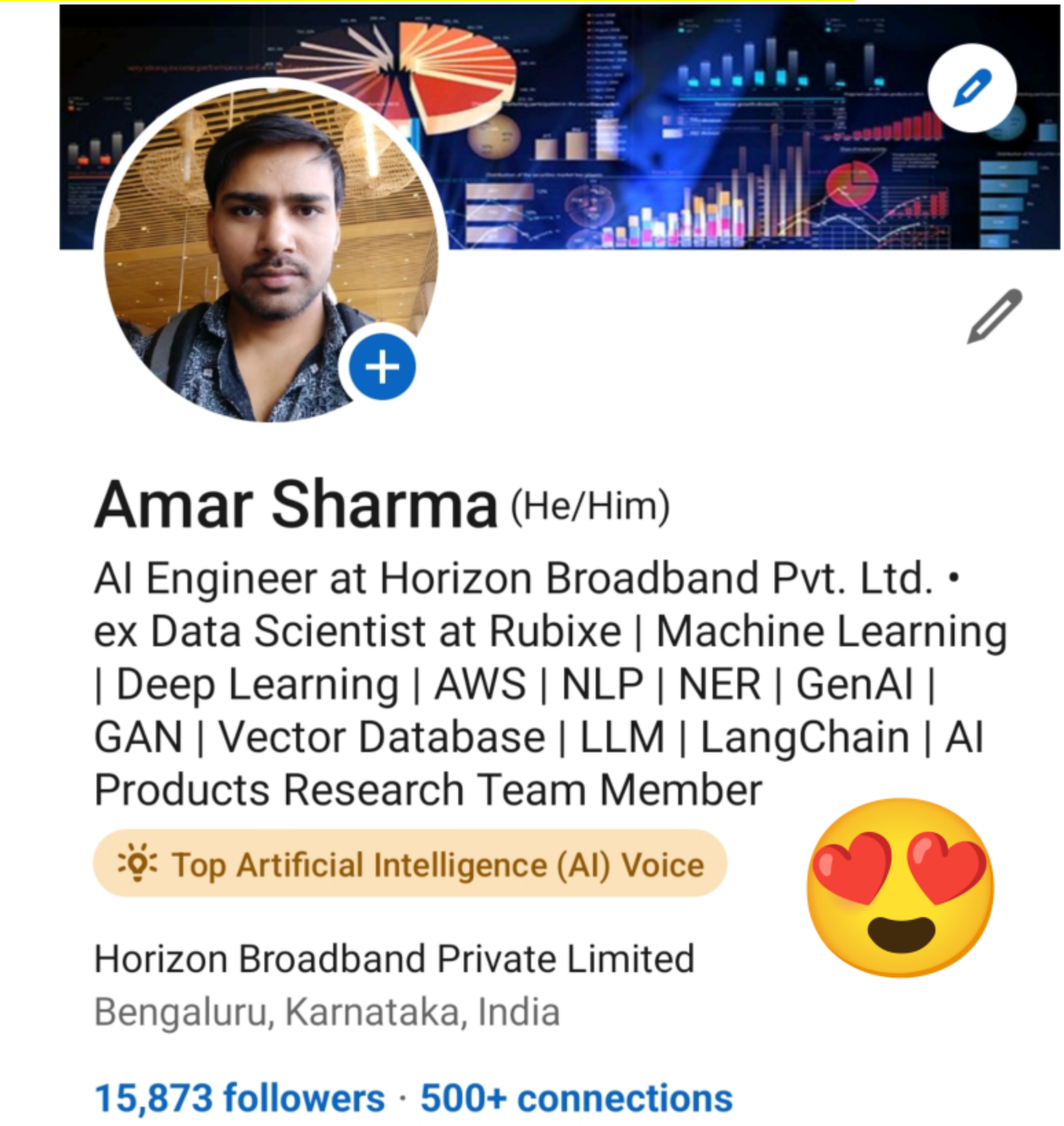
ROC Curve (Receiver Operating Characteristic):

- Plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- Useful when the classes are balanced.

Precision-Recall Curve:
- Plots precision against recall at various threshold settings.
- More informative than the ROC curve for

imbalanced datasets.

Follow for more informative content:



Amar Sharma (He/Him)

AI Engineer at Horizon Broadband Pvt. Ltd. • ex Data Scientist at Rubixe | Machine Learning | Deep Learning | AWS | NLP | NER | GenAI | GAN | Vector Database | LLM | LangChain | AI Products Research Team Member

Top Artificial Intelligence (AI) Voice

Horizon Broadband Private Limited

Bengaluru, Karnataka, India

15,873 followers · 500+ connections