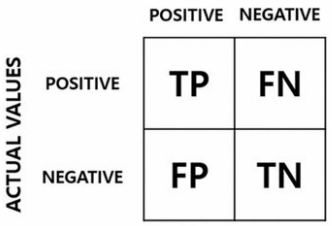
ML Viva Questions:

1. **A dataset is divided into 80-20, can you divide it into 50-50, if yes then what is it called?**

- Yes, a dataset can be divided into a 50-50 split. This is typically referred to as a train-test split, where 50% of the data is used for training, and 50% is used for testing, though the 80-20 split is more common in practice to retain more data for model training.

2. **Precision, Recall, F1-Score.**

- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives. It measures how accurate the positive predictions are.
- **Recall:** The ratio of correctly predicted positive observations to all actual positives. It shows how well the model identifies positive cases.
- **F1-Score:** The harmonic mean of Precision and Recall, balancing them into one metric that's useful when you have an uneven class distribution.



$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3. **Question on experiments.**

- *** VARIES EXPERIMENT TO EXPERIMENT. ***

4. **How do you evaluate a model?**

- Models are evaluated using performance metrics such as accuracy, precision, recall, F1-score, AUC-ROC for classification, and MSE or RMSE for regression. Cross-validation can also help validate results across different folds of data.

5. **Confusion matrix.**

- A confusion matrix is a table that summarizes the performance of a classification model by comparing predicted values with actual values. It includes four values: True Positives, False Positives, True Negatives, and False Negatives.

6. **Types of ml.**

- **Supervised Learning:** Models trained on labeled data (e.g., classification, regression).
- **Unsupervised Learning:** Models trained on unlabeled data (e.g., clustering, dimensionality reduction).
- **Reinforcement Learning:** Models learn by interacting with an environment and receiving feedback in the form of rewards.
- **Deep Learning, Deep Reinforcement Learning.**

7. **Make changes in the code (write).**

- Be ready to perform simple code edits or enhancements, such as modifying a train-test split, tuning hyperparameters, or adjusting a model function.

8. **Simple linear eq.**

- $y = \beta_0 + \beta_1 x$
- $\beta_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
- $\beta_0 = \bar{y} - \beta_1 \bar{x}$

9. **Mx + c, what is "c", why is it used?**

- M is the slope (change ing y for unit change in x)
- C is the intercept (value of y when x = 0)

10. **What is logistic regression?**

- Logistic regression is a classification algorithm used to predict binary outcomes. It outputs probabilities, which are converted into binary outcomes by applying a threshold (e.g., 0.5).

11. **What is CART?**

- CART (Classification and Regression Trees) is a decision tree algorithm used for both classification and regression tasks. It works by splitting data into subsets based on feature values.

12. **Bagging boosting.**

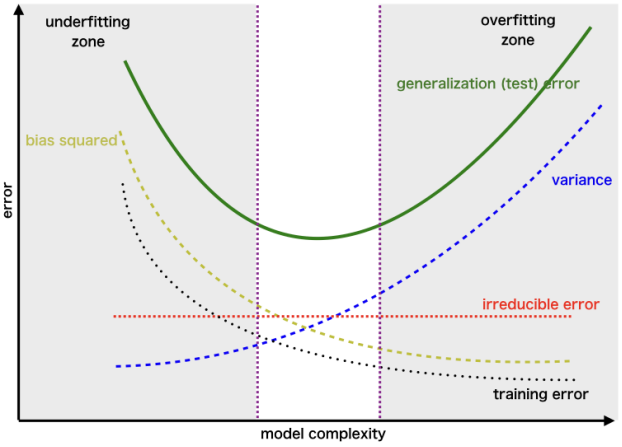
- **Bagging:** A technique that trains multiple instances of a model on random subsets of the dataset and averages their results (e.g., Random Forest).
- **Boosting:** A sequential technique where each model attempts to correct errors made by the previous ones (e.g., AdaBoost, Gradient Boosting).

13. What is ensemble learning?
- Ensemble learning combines predictions from multiple models to improve accuracy. Techniques include bagging, boosting, and stacking.

14. What is reinforcement learning?
- Reinforcement learning is a type of ML where agents learn by interacting with an environment, receiving rewards or penalties, and aiming to maximize cumulative rewards.

15. PCA.
- PCA is a dimensionality reduction technique that transforms features into a set of linearly uncorrelated components, helping to reduce the complexity of data.

16. Bias variance tradeoff.
- This is the balance between bias (error from overly simplistic models) and variance (error from complex models). A good model minimizes both to avoid underfitting and overfitting.



17. What is cross validation?
- Cross-validation is a technique to evaluate ML models by partitioning the data into multiple subsets (folds), training the model on some folds, and testing it on others. This helps improve model reliability.

18. What is knn?
- KNN is a simple, non-parametric algorithm used for classification and regression. It classifies a data point based on the majority label of its K nearest neighbors.

19. What is k fold?
- K-fold cross-validation splits data into K subsets (folds). The model is trained and tested K times, each time using a different fold as the test set and the remaining folds for training.

20. Example of unsupervised (real life).
- Customer segmentation in marketing is an example where clustering (an unsupervised learning technique) is used to group customers based on purchasing behavior without labeled data.