1. What is the definition of a target function? In the sense of a real-life example, express the target function. How is a target function's fitness assessed?

🡪 A target function, in machine learning, is a method for solving a problem that an AI algorithm parses its training data to find. The target function is essentially the formula that an algorithm feeds data to in order to calculate predictions.

Analysing the massive amounts of data related to its given problem, an AI derives understanding of previously unspecified rules by detecting consistencies in the data. The observations of inherent rules about how the studied subject operates inform the AI on how to process future data that does not include an output by applying this previously unknown function

2. What are predictive models, and how do they work? What are descriptive types, and how do you use them? Examples of both types of models should be provided. Distinguish between these two forms of models.

🡪 Predictive modelling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data.

It works by analysing current and historical data and projecting what it learns on a model generated to forecast likely outcomes.

The three main types of descriptive studies are Case studies, Naturalistic observation, and Surveys.

Some examples of descriptive research are: A specialty food group launching a new range of barbecue rubs would like to understand what Flavors of rubs are favoured by different people.

Case Studies are a type of observational research that involve a thorough descriptive analysis of a single individual, group, or event. There is no single way to conduct a case study so researchers use a range of methods from unstructured interviewing to direct observation.

3. Describe the method of assessing a classification model's efficiency in detail. Describe the various measurement parameters.

🡪 Assessing the efficiency of a classification model involves evaluating how well the model performs in categorizing instances into different classes. There are several measurement parameters that provide insights into different aspects of the model's performance. Here are the key measurement parameters commonly used to assess the efficiency of a classification model:

🡪Accuracy: Accuracy is a simple and widely used metric that measures the overall correctness of the model's predictions. It is the ratio of correctly predicted instances to the total instances in the dataset.

Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

🡪Precision: Precision focuses on the accuracy of the positive predictions made by the model. It answers the question: Of all instances predicted as positive, how many were actually positive?

Precision = (True Positives) / (True Positives + False Positives)

Precision is particularly useful when the cost of false positives is high (e.g., in medical diagnoses) and you want to minimize incorrect positive predictions.

🡪Recall (Sensitivity or True Positive Rate): Recall measures the ability of the model to correctly identify all relevant instances. It answers the question: Of all actual positive instances, how many were correctly predicted as positive?

Recall = (True Positives) / (True Positives + False Negatives)

Recall is crucial when missing actual positive instances is more concerning (e.g., in detecting fraud or diseases).

🡪F1-Score: The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure that takes into account both false positives and false negatives.

F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)

The F1-Score is useful when you want to find a balance between precision and recall.

🡪Specificity (True Negative Rate): Specificity measures the ability of the model to correctly identify negative instances.

Specificity = (True Negatives) / (True Negatives + False Positives)

Specificity is important in scenarios where correctly identifying negative instances is critical.

🡪 ROC Curve (Receiver Operating Characteristic Curve): The ROC curve is a graphical representation that illustrates the trade-off between the true positive rate (recall) and the false positive rate. AUC (Area Under the Curve) of the ROC curve quantifies the model's ability to distinguish between different classes.

🡪Confusion Matrix: A confusion matrix is a tabular representation that shows the number of true positives, true negatives, false positives, and false negatives. It's a useful way to visualize the performance of a classification model.

4.

i. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting?  
🡪In machine learning, underfitting refers to a situation where a model is too simple to capture the underlying patterns in the data, resulting in poor performance on both the training and testing/validation datasets. An underfit model fails to generalize well to new, unseen data points, leading to low accuracy and unsatisfactory predictive power.

The most common reason for underfitting is model complexity that is too low to capture the complexities present in the data.  
This often happens when:

-The Model is Too Simple

-Insufficient Training

-Limited Features

-Over-regularization

-Small Training Dataset

-Ignoring Important Features

-Data Noise

ii. What does it mean to overfit? When is it going to happen?  
🡪 Overfitting occurs when a model learns the training data too well, to the point that it captures the noise and random fluctuations in the data rather than the underlying patterns. As a result, an overfit model performs exceptionally well on the training data but generalizes poorly to new, unseen data. Overfitting can lead to reduced predictive power and poor performance on validation or test datasets.

Overfitting is more likely to happen when:

-Model Complexity is Too High

-Insufficient Training Data

-Too Many Features

-Overemphasis on Noise

-Not Enough Regularization

-Overfitting in Deep Learning

iii. In the sense of model fitting, explain the bias-variance trade-off.  
🡪 The bias-variance trade-off is a fundamental concept in machine learning that highlights the relationship between two sources of error that affect a model's performance: bias and variance. Achieving a balance between bias and variance is crucial for creating models that generalize well to new, unseen data.

-Bias

-Variance

-High Bias, Low Variance

-Low Bias, High Variance

-Balanced Bias and Variance

5. Is it possible to boost the efficiency of a learning model? If so, please clarify how.  
🡪 Building a machine learning model is not enough to get the right predictions, as you have to check the accuracy and need to validate the same to ensure get the precise results. And validating the model will improve the performance of the ML model. Some ways of boosting the efficiency of a learning model are mentioned below:

Add more Data Samples

Look at the problem differently: Looking at the problem from a new perspective can add valuable information to your model and help you uncover hidden relationships between the story variables. Asking different questions may lead to better results and, eventually, better accuracy.

Adding Context to Data: More context can always lead to a better understanding of the problem and, eventually, better performance of the model. Imagine we are selling a car, a BMW. That alone doesn’t give us much information about the car. But, if we add the color, model and distance travelled, then you’ll start to have a better picture of the car and its possible value.

Finetuning our hyperparameter: to get the answer, we will need to do some trial and error until you reach your answer.

Train our model using cross-validation

Experimenting with different Algorithms.

6. How would you rate an unsupervised learning model's success? What are the most common success indicators for an unsupervised learning model?  
🡪 In case of supervised learning, it is mostly done by measuring the performance metrics such as accuracy, precision, recall, AUC, etc. on the training set and the holdout sets whereas for Unsupervised Learning it is different. Since there is no pre-evidence or records for patterns, we cannot directly compute the accuracy by comparing actual and predicted outputs but there exist many evaluation metrics to measure the performance of unsupervised learning algorithms after the training process.

Some of them are-

Clustering - Jaccard similarity index, Rand Index, Purity, Silhouette measure, Sum of squared errors, etc.

Association rule mining – Lift, Confidence

Time series analysis – Root mean square error, mean absolute error, mean absolute percentage error, etc.

Autoencoders - Reconstruction errors

Natural Language processing (like sentiment analysis and text clustering) – Comparing the correlation between natural words after converting them to numerical vectors.

Principal component analysis – Reconstruction error, Scree plot

Generative adversarial networks – Discriminator functions

Recurrent neural networks and LSTM (In numerical series) – Root mean square error, mean absolute error, mean absolute percentage error, etc.

Recurrent neural networks and LSTM (In semantic series) - Word to vector correlation

Anomaly detection (like DBSCAN, OPTICS) – Cohesion, Separation, Sum of squared errors, etc.

Expectation/ Maximization problems – Log-likelihood

Survival analysis (Cox model 1) – Simple hazard ration, R Squared

Survival analysis (Cox model 2) – Two group hazard ratio and brier score, Log-rank test, Somers’ rank correlation, Time-dependent ROC – AUC, Power validation, etc.

7. Is it possible to use a classification model for numerical data or a regression model for categorical data with a classification model? Explain your answer.  
🡪 Categorical Data is the data that generally takes a limited number of possible values. Also, the data in the category need not be numerical, it can be textual in nature. All machine learning models are some kind of mathematical model that need numbers to work with. This is one of the primary reasons we need to pre-process the categorical data before we can feed it to machine learning models.

If a categorical target variable needs to be encoded for a classification predictive modelling problem, then the Label Encoder class can be used.

8. Describe the predictive modeling method for numerical values. What distinguishes it from categorical predictive modeling?  
🡪 predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analysing current and historical data and projecting what it learns on a model generated to forecast likely outcomes.

Classification is the process of identifying the category or class label of the new observation to which it belongs. Predication is the process of identifying the missing or unavailable numerical data for a new observation. That is the key difference between classification and prediction.

9. The following data were collected when using a classification model to predict the malignancy of a group of patients' tumors:

i. Accurate estimates – 15 cancerous, 75 benign

ii. Wrong predictions – 3 cancerous, 7 benign

Determine the model's error rate, Kappa value, sensitivity, precision, and F-measure.

🡪- Accurate estimates:

- True Positives (TP) = 15 (cancerous correctly predicted)

- True Negatives (TN) = 75 (benign correctly predicted)

- Wrong predictions:

- False Positives (FP) = 7 (benign incorrectly predicted as cancerous)

- False Negatives (FN) = 3 (cancerous incorrectly predicted as benign)

Total number of samples = TP + TN + FP + FN = 15 + 75 + 7 + 3 = 100

1. Error Rate:

Error Rate = (FP + FN) / Total Number of Samples

Error Rate = (7 + 3) / 100 = 0.1 or 10%

2. Kappa Value:

Kappa Value is a measure of agreement between the actual classifications and the model's predictions, considering the agreement that could occur by chance.

Kappa Value = (Accuracy - Chance Agreement) / (1 - Chance Agreement)

First, calculate the Accuracy:

Accuracy = (TP + TN) / Total Number of Samples = (15 + 75) / 100 = 0.9 or 90%

Next, calculate the Chance Agreement:

Chance Agreement = (TP + FP) \* (TP + FN) / Total Number of Samples^2 + (TN + FP) \* (TN + FN) / Total Number of Samples^2

Chance Agreement = (15 + 7) \* (15 + 3) / 100^2 + (75 + 7) \* (75 + 3) / 100^2

Chance Agreement = 0.22

Now calculate the Kappa Value:

Kappa Value = (0.9 - 0.22) / (1 - 0.22) = 0.755

3. Sensitivity (True Positive Rate):

Sensitivity = TP / (TP + FN) = 15 / (15 + 3) = 0.8333 or 83.33%

4. Precision:

Precision = TP / (TP + FP) = 15 / (15 + 7) = 0.6818 or 68.18%

5. F-Measure (F1-Score):

F1-Score = 2 \* (Precision \* Sensitivity) / (Precision + Sensitivity)

F1-Score = 2 \* (0.6818 \* 0.8333) / (0.6818 + 0.8333) = 0.7504 or 75.04%

To summarize:

- Error Rate: 10%

- Kappa Value: 0.755

- Sensitivity: 83.33%

- Precision: 68.18%

- F-Measure: 75.04%  
  
10. Make quick notes on:

1. The process of holding out  
🡪 The hold-out method for training machine learning model is the process of splitting the data in different splits and using one split for training the model and other splits for validating and testing the models. The hold-out method is used for both model evaluation and model selection.

2. Cross-validation by tenfold  
🡪 10-fold cross validation would perform the fitting procedure a total of ten times, with each fit being performed on a training set consisting of 90% of the total training set selected at random, with the remaining 10% used as a hold out set for validation.

3. Adjusting the parameters  
🡪 A fancy name for training: the selection of parameter values, which are optimal in some desired sense (eg. minimize an objective function you choose over a dataset you choose). The parameters are the weights and biases of the network

11. Define the following terms:

1. Purity vs. Silhouette width  
🡪 Purity is a measure of the extent to which clusters contain a single class. Its calculation can be thought of as follows: For each cluster, count the number of data points from the most common class in said cluster.

The silhouette width is also an estimate of the average distance between clusters. Its value is comprised between 1 and -1 with a value of 1 indicating a very good cluster.

2. Boosting vs. Bagging  
🡪 Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data.

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

3. The eager learner vs. the lazy learner  
🡪 A lazy learner delays abstracting from the data until it is asked to make a prediction.

while an eager learner abstracts away from the data during training and uses this abstraction to make predictions rather than directly compare queries with instances in the dataset.