**Assignment 5**

Naive Approach:

1. What is the Naive Approach in machine learning?

Ans: The Naïve Bayes classifier is a probabilistic supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms. Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem.

1. Explain the assumptions of feature independence in the Naive Approach.

Ans: The assumptions of feature independence means that combinations of features are treated in just the same way as each feature appearing on its own.By doing so, the joint distribution can be found easily by just multiplying the probability of each feature whilst in the real world they may not be independent and you have to find the correct joint distribution. It is naive due to this simplification. For example, the temperature being ‘Hot’ has nothing to do with the humidity or the outlook being ‘Rainy’ has no effect on the winds. Hence, the features are assumed to be independent.

1. How does the Naive Approach handle missing values in the data?

Ans: Naive Bayes apparently handles missing data differently, depending on whether they exist in training or testing/classification instances.When classifying instances, the attribute with the missing value is simply not included in the probability calculation.In training, the instance [with the missing data] is not included in frequency count for attribute value-class combination

1. What are the advantages and disadvantages of the Naive Approach?

Ans: Naive Approach:

Advantages :-

When the independent assumption holds then this classifier gives outstanding accuracy. Easy to implement as only the probability is to be calculated

It works well with high dimensions such as text classification.

Disadvantages :- If the independent assumption does not hold then performance is very low

1. Can the Naive Approach be used for regression problems? If yes, how?

Ans:Naïve Approach is commonly used for classification, so it **can** be either 1 or 0 nothing in between like 0.5 (regression) Even if we force naive bayes and tweak it a little bit **for** regression the result is disappointing

1. How do you handle categorical features in the Naive Approach?

Ans: The way to deal with categorical data is to create each category as a feature and with boolean values.

.

1. What is Laplace smoothing and why is it used in the Naive Approach?

Ans: Laplace smoothing is a smoothing technique that helps tackle the problem of zero probability in the Nave Bayes machine learning algorithm. Using higher alpha(smoothing parameter) values will push the likelihood towards a value of 0.5, i.e., the probability of a word equal to 0.5 for both the positive and negative reviews. If we choose a value of alpha!=0 (not equal to 0), the probability will no longer be zero even if a word is not present in the training dataset.

1. How do you choose the appropriate probability threshold in the Naive Approach?

Ans: Naive Bayes has the tendency to push probabilities to extremes (0% and 100%). set the thresholds as a value in the range [0, 1], such that they sum to 1. This will get you the desired rule of "Classify as True if the probability is over threshold T, otherwise classify as False".

1. Give an example scenario where the Naive Approach can be applied.

Ans: weather conditions for playing, pattern searching

KNN:

1. What is the K-Nearest Neighbors (KNN) algorithm?

Ans:  k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

1. How does the KNN algorithm work?

### Ans [KNN algorithm is a **supervised learning classifier** that uses proximity or distance to assign a class label to a new data point based on the labels of its nearest neighbors](https://www.bing.com/ck/a?!&&p=6757f2716c5b3e04JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc2OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+KNN+algorithm+work%3f%0d%0a&u=a1aHR0cHM6Ly93d3cuaWJtLmNvbS90b3BpY3Mva25u&ntb=1). [It is a non-parametric algorithm, meaning it does not make any assumptions on the data distribution](https://www.bing.com/ck/a?!&&p=5243768f249a9d93JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc3Mg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+KNN+algorithm+work%3f%0d%0a&u=a1aHR0cHM6Ly93d3cubXlncmVhdGxlYXJuaW5nLmNvbS9ibG9nL2tubi1hbGdvcml0aG0taW50cm9kdWN0aW9uLw&ntb=1)[3](https://www.bing.com/ck/a?!&&p=4f214bc578ae6c0dJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc3Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+KNN+algorithm+work%3f%0d%0a&u=a1aHR0cHM6Ly93d3cubXlncmVhdGxlYXJuaW5nLmNvbS9ibG9nL2tubi1hbGdvcml0aG0taW50cm9kdWN0aW9uLw&ntb=1). [It is also a lazy learning algorithm, meaning it defers the computation until classification](https://www.bing.com/ck/a?!&&p=2817b6d142c8ba61JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc3NA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+KNN+algorithm+work%3f%0d%0a&u=a1aHR0cHM6Ly93d3cuaWJtLmNvbS9kb2NzL2VuL2lhcz90b3BpYz1rbm4tdXNhZ2U&ntb=1). [It works on the principle of information gain, finding the most suitable neighbors to predict an unknown value](https://www.bing.com/ck/a?!&&p=38208ad91c68e57cJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc3Ng&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+KNN+algorithm+work%3f%0d%0a&u=a1aHR0cHM6Ly9uZXB0dW5lLmFpL2Jsb2cva25uLWFsZ29yaXRobS1leHBsYW5hdGlvbi1vcHBvcnR1bml0aWVzLWxpbWl0YXRpb25z&ntb=1)[**5**](https://www.bing.com/ck/a?!&&p=67250f2de6fa03faJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc3Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+the+KNN+algorithm+work%3f%0d%0a&u=a1aHR0cHM6Ly9uZXB0dW5lLmFpL2Jsb2cva25uLWFsZ29yaXRobS1leHBsYW5hdGlvbi1vcHBvcnR1bml0aWVzLWxpbWl0YXRpb25z&ntb=1)

1. How do you choose the value of K in KNN?

Ans: The value of k is very crucial in the KNN algorithm to define the number of neighbors in the algorithm. The value of k in the k-nearest neighbors (k-NN) algorithm should be chosen based on the input data. If the input data has more outliers or noise, a higher value of k would be better. It is recommended to choose an odd value for k to avoid ties in classification. [Cross-validation](https://www.geeksforgeeks.org/cross-validation-machine-learning/) methods can help in selecting the best k value for the given dataset. Normally, chose the value of k is k = sqrt(N)/2, where N stands for the number of samples in your training dataset. Most of the time below approach is followed in industry. Initialize a random K value and start computing. Derive a plot between error rate and K denoting values in a defined range. Then choose the K value as having a minimum error rate. Derive a plot between accuracy and K denoting values in a defined range. Then choose the K value as having a maximum accuracy

1. What are the advantages and disadvantages of the KNN algorithm?.

Ans: Advantages of K-NN Algorithm It is simple to implement. No training is required before classification. Disadvantages of K-NN Algorithm Can be cost-intensive when working with a large data set. A lot of memory is required for processing large data sets. Choosing the right value of K can be tricky.

1. How does the choice of distance metric affect the performance of KNN?

Ans: In KNN classifier, the distances between the test sample and the training data samples are identified by different measures. A nonconvex distance performed the best when applied on most data sets comparing with the other tested distances.

1. Can KNN handle imbalanced datasets? If yes, how?

Ans:yes, a neighborhood based under-sampling technique to handle class imbalance nature of classification dataset or use SMOTE

1. How do you handle categorical features in KNN?

Ans: [KNN doesn't handle categorical features well, which is a fundamental weakness of the algorithm](https://www.bing.com/ck/a?!&&p=0c1d8c635f15bfb5JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTU5OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+handle+categorical+features+in+KNN%3f%0d%0a&u=a1aHR0cHM6Ly9kYXRhc2NpZW5jZS5zdGFja2V4Y2hhbmdlLmNvbS9xdWVzdGlvbnMvMjY3MTMvaG93LWRvZXMta25uLWhhbmRsZS1jYXRlZ29yaWNhbC1mZWF0dXJlcw&ntb=1). [KNN doesn't work well when features are on different scales, especially when one of the 'scales' is a category label](https://www.bing.com/ck/a?!&&p=14ba9a53e0b2555eJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYwMA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+handle+categorical+features+in+KNN%3f%0d%0a&u=a1aHR0cHM6Ly9kYXRhc2NpZW5jZS5zdGFja2V4Y2hhbmdlLmNvbS9xdWVzdGlvbnMvMjY3MTMvaG93LWRvZXMta25uLWhhbmRsZS1jYXRlZ29yaWNhbC1mZWF0dXJlcw&ntb=1). [Among the three classification methods, only Kernel Density Classification can handle categorical variables in theory, while kNN and SVM are unable to be applied directly since they are based on the Euclidean distances](https://www.bing.com/ck/a?!&&p=51e22ed88e597de0JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYwMg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+handle+categorical+features+in+KNN%3f%0d%0a&u=a1aHR0cHM6Ly9zdGF0cy5saWJyZXRleHRzLm9yZy9Cb29rc2hlbHZlcy9Db21wdXRpbmdfYW5kX01vZGVsaW5nL1JURyUyNTNBX0NsYXNzaWZpY2F0aW9uX01ldGhvZHMvNCUyNTNBX051bWVyaWNhbF9FeHBlcmltZW50c19hbmRfUmVhbF9EYXRhX0FuYWx5c2lzL1ByZXByb2Nlc3Npbmdfb2ZfY2F0ZWdvcmljYWxfcHJlZGljdG9yc19pbl9TVk0lMjUyQ19LTk5fYW5kX0tEQ18oY29udHJpYnV0ZWRfYnlfWGlfQ2hlbmcp&ntb=1). [To handle categorical variables in KNN, create dummy variables out of a categorical variable and include them instead of the original categorical](https://www.bing.com/ck/a?!&&p=3e7278d038bd4a2bJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTYwNA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+handle+categorical+features+in+KNN%3f%0d%0a&u=a1aHR0cHM6Ly93d3cubGlzdGVuZGF0YS5jb20vMjAxNy8xMi9rLW5lYXJlc3QtbmVpZ2hib3Itc3RlcC1ieS1zdGVwLXR1dG9yaWFsLmh0bWw&ntb=1)

1. What are some techniques for improving the efficiency of KNN?

Ans:Techniques for improving the efficiency of KNN are : use unigrams, TF-IDF vectors, try including stop words, use both unigrams and bigrams instead of only bigrams as the latter might be overfitting,use dimensionality reduction techniques.

1. Give an example scenario where KNN can be applied.

Ans: Text mining,Agriculture,Finance,Medical,Facial recognition

Clustering:

1. What is clustering in machine learning?

Ans: **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

1. Explain the difference between hierarchical clustering and k-means clustering.

* Ans:
* In K-means clustering, the number of clusters is pre-defined and is denoted by “K”, but in hierarchical clustering, the number of sets is either one or similar to the number of data observations.
* K-means algorithm in all its iterations has the same number of clusters, while hierarchical clustering is a purely agglomerative approach and goes on to build one giant cluster.
* K-means need circular data, while hierarchical clustering has no such requirement.
* In hierarchical clustering, results are reproducible, while K-means clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. A hierarchical clustering is a set of nested clusters that are arranged as a tree.

1. How do you determine the optimal number of clusters in k-means clustering?

Ans: Elbow method is used to determine the optimal number of clusters in k-means clustering.steps are:

* Compute clustering algorithm (e.g., k-means clustering) for different values of k.
* For each k, calculate the total within-cluster sum of square (wss).
* Plot the curve of wss according to the number of clusters k.
* The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.
* Compute the average WSS for 20 runs of k-means on an increasing number of clusters (starting with 2, and ending with say 9 or 10), and keep the solution that has minimal WSS over this clusters set.

1. What are some common distance metrics used in clustering?

Ans:Euclidean distantce,Manhanthan distance,Minkowaski distance

1. How do you handle categorical features in clustering?

Ans: Use following techniques for handling categorical features in clustering

* One Hot Encoding:use for nominal categorical , In one Hot Encoding method, each category value is converted into a new column and assigned a value as 1 or 0 to the column.
* One Hot Encoding with multiple categories:  instead of creating the new column for every category, they limit creating the new column for 10 most frequent categories.
* Ordinal Number Encoding: each unique category value is given an integer value
* Count or Frequency Encoding: substitute the categories by the count of the observations that show that category in the dataset.
* Target guided Ordinal Encoding: transform our categorical variable by comparing it to the target or output variable.
* Mean Ordinal Encoding: It’s a sight variant of target-guided ordinal encoding , the category is replaced with the obtained mean value

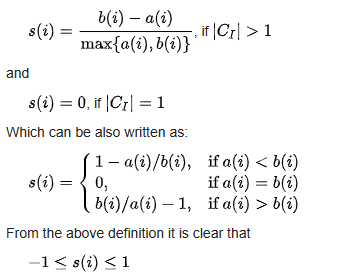
· Probability Ratio Encoding: This technique is suitable for classification problems only when the target variable is binary(Either 1 or 0 or True or False).In this technique, we will substitute the category value with the probability ratio i.e. P(1)/P(0).

1. What are the advantages and disadvantages of hierarchical clustering?

Ans: Advantages:Robustness, Easy to interpret,Flexible, Scalable Greater accuracy

Disadvantages: It is sensitive to outliers, Hierarchical clustering is computationally expensive, Hierarchical clustering is computationally expensive, Agglomerative or divisive clustering is prone to producing overlapping clusters, Hierarchical clustering methods require a predetermined number of clusters before they can begin clustering.

1. Explain the concept of silhouette score and its interpretation in clustering.

Ans: The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. 

For data point �∈�� (data point i in the cluster ��),The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.

26. Give an example scenario where clustering can be applied.

Ans:

1. Few Applications Customer Service: Clustering is used to group customer inquiries and complaints into categories, identify common issues, and develop targeted solutions.
2. Manufacturing: Clustering is used to group similar products together, optimize production processes, and identify defects in manufacturing processes.
3. Medical diagnosis: Clustering is used to group patients with similar symptoms or diseases, which helps in making accurate diagnoses and identifying effective treatments.
4. Fraud detection: Clustering is used to identify suspicious patterns or anomalies in financial transactions, which can help in detecting fraud or other financial crimes.
5. Traffic analysis: Clustering is used to group similar patterns of traffic data, such as peak hours, routes, and speeds, which can help in improving transportation planning and infrastructure.

Anomaly Detection:

27. What is anomaly detection in machine learning?

Ans: **Anomaly Detection** is the technique of identifying rare events /observations/data points which can raise suspicions by being statistically different from the rest of the observations.

28. Explain the difference between supervised and unsupervised anomaly detection.

Ans:

Supervised Anomaly Detection: This method requires a labeled dataset containing both normal and anomalous samples to construct a predictive model to classify future data points. The most commonly used algorithms for this purpose are supervised Neural Networks, [Support Vector Machine learning](https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-python/), [K-Nearest Neighbors Classifier](https://www.geeksforgeeks.org/k-nearest-neighbours/), etc.

Unsupervised Anomaly Detection: This method does require any training data and instead assumes two things about the data ie Only a small percentage of data is anomalous and Any anomaly is statistically different from the normal samples. Based on the above assumptions, the data is then clustered using a similarity measure and the data points which are far off from the cluster are considered to be anomalies.

29. What are some common techniques used for anomaly detection?

Ans: Common Techniques:

* Density-based techniques (k-nearest neighbor, local outlier factor, isolation forests, and many more variations of this concept)
* Subspace-, correlation-based and tensor-based outlier detection for high-dimensional data
* Neural Networks-Based
* Bayesian Networks Based
* Support Vector Machines Based

30. How does the One-Class SVM algorithm work for anomaly detection?

Ans: [One-Class Support Vector Machine (SVM) is an unsupervised model for anomaly or outlier detection](https://www.bing.com/ck/a?!&&p=26c92ac07ea17574JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1Ng&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+the+One-Class+SVM+algorithm+work+for+anomaly+detection%3f%0d%0a&u=a1aHR0cHM6Ly9tZWRpdW0uY29tL2dyYWJuZ29pbmZvL29uZS1jbGFzcy1zdm0tZm9yLWFub21hbHktZGV0ZWN0aW9uLTZjOTdmZGQ2ZDhhZg&ntb=1). [It learns the boundary for the normal data points and identifies the data outside the border to be anomalies](https://www.bing.com/ck/a?!&&p=48313e881b747041JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+the+One-Class+SVM+algorithm+work+for+anomaly+detection%3f%0d%0a&u=a1aHR0cHM6Ly9tZWRpdW0uY29tL2dyYWJuZ29pbmZvL29uZS1jbGFzcy1zdm0tZm9yLWFub21hbHktZGV0ZWN0aW9uLTZjOTdmZGQ2ZDhhZg&ntb=1). [OneClassSVM is a model object used for outlier detection and novelty detection](https://www.bing.com/ck/a?!&&p=56e93aba37626fbbJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2MA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+the+One-Class+SVM+algorithm+work+for+anomaly+detection%3f%0d%0a&u=a1aHR0cHM6Ly93d3cubWF0aHdvcmtzLmNvbS9oZWxwL3N0YXRzL29uZWNsYXNzc3ZtLmh0bWw&ntb=1" \t "_blank). [The ocsvm function trains a OneClassSVM object and returns anomaly indicators and scores for the training data](https://www.bing.com/ck/a?!&&p=8b185d5b6ad1f2eaJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2Mg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+the+One-Class+SVM+algorithm+work+for+anomaly+detection%3f%0d%0a&u=a1aHR0cHM6Ly93d3cubWF0aHdvcmtzLmNvbS9oZWxwL3N0YXRzL29uZWNsYXNzc3ZtLmh0bWw&ntb=1). [The One-class SVM applies a One-class classification method for novelty detection](https://www.bing.com/ck/a?!&&p=7409713eabc3ce49JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2NA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+the+One-Class+SVM+algorithm+work+for+anomaly+detection%3f%0d%0a&u=a1aHR0cHM6Ly93d3cuZGF0YXRlY2hub3Rlcy5jb20vMjAyMC8wNC9hbm9tYWx5LWRldGVjdGlvbi13aXRoLW9uZS1jbGFzcy1zdm0uaHRtbA&ntb=1)[3](https://www.bing.com/ck/a?!&&p=5c15f54f611af88cJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+the+One-Class+SVM+algorithm+work+for+anomaly+detection%3f%0d%0a&u=a1aHR0cHM6Ly93d3cuZGF0YXRlY2hub3Rlcy5jb20vMjAyMC8wNC9hbm9tYWx5LWRldGVjdGlvbi13aXRoLW9uZS1jbGFzcy1zdm0uaHRtbA&ntb=1). [One-Class Support Vector Machines (OCSVM) are one of the state-of-the-art approaches for novelty detection in machine learning, due to their flexibility in fitting complex nonlinear boun](https://www.bing.com/ck/a?!&&p=be64b95cc351b6a5JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2Ng&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+the+One-Class+SVM+algorithm+work+for+anomaly+detection%3f%0d%0a&u=a1aHR0cHM6Ly9hcnhpdi5vcmcvYWJzLzIxMDQuMTExNDY&ntb=1)daries between novel data and noramal.

31. How do you choose the appropriate threshold for anomaly detection?

Ans:

* Construct a train set using a large sample of observations without anomalies.
* Take a smaller sample of observations containing anomalies (manually labelled) and use it to construct a validation and test set.
* Now train your anomaly detection model while tuning the threshold using the validation set and additionally using the test set to evaluate your model.

32. How do you handle imbalanced datasets in anomaly detection?

Ans: Techniques:

* Ensemble algorithms based on robot vision that combine multiple semi-supervised k-means methods and C4.5 to relieve the problems of imbalanced data in anomaly detection.
* Anomaly detection aided budget online weighted learning method (BOW-LM) based on the widely used Feedforward Networks with Random Weights.
* One-Class Classification (OCC) that involves fitting a model on the “normal” data and predicting whether new data is normal or an outlier/anomaly.

33. Give an example scenario where anomaly detection can be applied.

Ans: Examples:

* Sudden burst or decrease in activity
* Error in the text
* Sudden rapid drop or increase in temperature
* Adding custom URLs to machine learning results
* Aggregating data for faster performance

Dimension Reduction:

34. What is dimension reduction in machine learning?

Ans: [Dimension reduction in machine learning is the process of reducing the number of features in a dataset while retaining as much information as possible](https://www.bing.com/ck/a?!&&p=28a43f74e2bcd60eJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcyMQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=+What+is+dimension+reduction+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cuZ2Vla3Nmb3JnZWVrcy5vcmcvZGltZW5zaW9uYWxpdHktcmVkdWN0aW9uLw&ntb=1). [It can be used to reduce the complexity of a model, improve the performance of a learning algorithm, or make it easier to visualize the data](https://www.bing.com/ck/a?!&&p=77fa9a431dff252eJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcyNA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=+What+is+dimension+reduction+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cuZ2Vla3Nmb3JnZWVrcy5vcmcvZGltZW5zaW9uYWxpdHktcmVkdWN0aW9uLw&ntb=1). [Dimension reduction techniques include feature selection, PCA, linear algebra methods, projection methods, and autoencoders](https://www.bing.com/ck/a?!&&p=cb36ce86b55f6defJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcyOA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=+What+is+dimension+reduction+in+machine+learning%3f&u=a1aHR0cHM6Ly9tYWNoaW5lbGVhcm5pbmdtYXN0ZXJ5LmNvbS9kaW1lbnNpb25hbGl0eS1yZWR1Y3Rpb24tZm9yLW1hY2hpbmUtbGVhcm5pbmcv&ntb=1) etc.

35. Explain the difference between feature selection and feature extraction.

Ans: [The key difference between feature selection and feature extraction is that feature selection keeps a subset of the original features while feature extraction algorithms transform the data onto a new feature space](https://www.bing.com/ck/a?!&&p=9354860cc0bac450JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+difference+between+feature+selection+and+feature+extraction.&u=a1aHR0cHM6Ly9oMm8uYWkvd2lraS9mZWF0dXJlLXNlbGVjdGlvbi8&ntb=1). [Feature selection involves choosing a subset of the original pool of features](https://www.bing.com/ck/a?!&&p=c291bc92ec46f99aJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwNQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+difference+between+feature+selection+and+feature+extraction.&u=a1aHR0cHM6Ly9xdWFudGRhcmUuY29tL3doYXQtaXMtdGhlLWRpZmZlcmVuY2UtYmV0d2Vlbi1mZWF0dXJlLWV4dHJhY3Rpb24tYW5kLWZlYXR1cmUtc2VsZWN0aW9uLw&ntb=1). [Feature extraction, on the other hand, involves getting useful features from existing data](https://www.bing.com/ck/a?!&&p=daab3cd7c3ed6b2bJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwNw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+difference+between+feature+selection+and+feature+extraction.&u=a1aHR0cHM6Ly9xdWFudGRhcmUuY29tL3doYXQtaXMtdGhlLWRpZmZlcmVuY2UtYmV0d2Vlbi1mZWF0dXJlLWV4dHJhY3Rpb24tYW5kLWZlYXR1cmUtc2VsZWN0aW9uLw&ntb=1). [Examples of feature extraction include extraction of contours in images, extraction of digrams from a text, extraction](https://www.bing.com/ck/a?!&&p=e0067bfa6b857cc6JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcxMA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+difference+between+feature+selection+and+feature+extraction.&u=a1aHR0cHM6Ly9kYXRhc2NpZW5jZS5zdGFja2V4Y2hhbmdlLmNvbS9xdWVzdGlvbnMvMTMwL3doYXQtaXMtZGltZW5zaW9uYWxpdHktcmVkdWN0aW9uLXdoYXQtaXMtdGhlLWRpZmZlcmVuY2UtYmV0d2Vlbi1mZWF0dXJlLXNlbGVjdGk&ntb=1)phonemes from recording of spoken text.

36. How does Principal Component Analysis (PCA) work for dimension reduction?

Ans: [Principal Component Analysis (PCA) is a dimensionality-reduction method that transforms a large set of variables into a smaller one that still contains most of the information in the large set](https://www.bing.com/ck/a?!&&p=0c035d18a75d6649JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+Principal+Component+Analysis+(PCA)+work+for+dimension+reduction%3f&u=a1aHR0cHM6Ly9idWlsdGluLmNvbS9kYXRhLXNjaWVuY2Uvc3RlcC1zdGVwLWV4cGxhbmF0aW9uLXByaW5jaXBhbC1jb21wb25lbnQtYW5hbHlzaXM&ntb=1) and the co-relation between the newly obtained Principal Components is minimum. [PCA works by finding the lower-dimensional surface to project the high-dimensional data, while preserving as much of the data's variation as possible](https://www.bing.com/ck/a?!&&p=bdc8255785723f57JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+does+Principal+Component+Analysis+(PCA)+work+for+dimension+reduction%3f&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvUHJpbmNpcGFsX2NvbXBvbmVudF9hbmFseXNpcw&ntb=1). The new features obtained after applying PCA are called Principal Components and are denoted as PCi (i=1,2,3…n). Here, (Principal Component-1) PC1 captures the maximum information of the original dataset, followed by PC2, then PC3 and so on.

37. How do you choose the number of components in PCA?

Ans: steps:

* [Select the first two or three principal components for visualization purposes](https://www.bing.com/ck/a?!&&p=b1b082b197f52ba2JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+choose+the+number+of+components+in+PCA%3fgeeks+for+geeks&u=a1aHR0cHM6Ly9zdGF0aXN0aWNzZ2xvYmUuY29tL2Nob29zZS1vcHRpbWFsLW51bWJlci1jb21wb25lbnRzLXBjYQ&ntb=1)[1](https://www.bing.com/ck/a?!&&p=bca9a32ab91674c8JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3Mg&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+choose+the+number+of+components+in+PCA%3fgeeks+for+geeks&u=a1aHR0cHM6Ly9zdGF0aXN0aWNzZ2xvYmUuY29tL2Nob29zZS1vcHRpbWFsLW51bWJlci1jb21wb25lbnRzLXBjYQ&ntb=1).
* [Choose a subset of principal components that can explain at least 90% of the variance in the data](https://www.bing.com/ck/a?!&&p=22307cb1c71e7062JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+choose+the+number+of+components+in+PCA%3fgeeks+for+geeks&u=a1aHR0cHM6Ly9zdGFja292ZXJmbG93LmNvbS9xdWVzdGlvbnMvNDY1MTY0NjkvcHJpbmNpcGFsLWNvbXBvbmVudC1hbmFseXNpcy1ob3ctbWFueS1jb21wb25lbnRz&ntb=1).
* [Plot the cumulative explained variance as a function of the number of components and look for an elbow point](https://www.bing.com/ck/a?!&&p=f8c1ce68899f8527JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+choose+the+number+of+components+in+PCA%3fgeeks+for+geeks&u=a1aHR0cHM6Ly9zdGFja292ZXJmbG93LmNvbS9xdWVzdGlvbnMvNTM4MDIwOTgvaG93LXRvLWNob29zZS10aGUtbnVtYmVyLW9mLWNvbXBvbmVudHMtcGNhLXNjaWtpdGxpZWFy&ntb=1).
* [Use the option that allows you to set the desired explained variance and let the algorithm determine the optimal number of components](https://www.bing.com/ck/a?!&&p=90a74d07d9b6726dJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+choose+the+number+of+components+in+PCA%3fgeeks+for+geeks&u=a1aHR0cHM6Ly93d3cubWlrdWxza2liYXJ0b3N6Lm5hbWUvcGNhLWhvdy10by1jaG9vc2UtdGhlLW51bWJlci1vZi1jb21wb25lbnRzLw&ntb=1).
* [Scale the data to the range between 0 and 1 before applying PCA](https://www.bing.com/ck/a?!&&p=9e4a4ad9d2839284JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY3OQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=How+do+you+choose+the+number+of+components+in+PCA%3fgeeks+for+geeks&u=a1aHR0cHM6Ly93d3cubWlrdWxza2liYXJ0b3N6Lm5hbWUvcGNhLWhvdy10by1jaG9vc2UtdGhlLW51bWJlci1vZi1jb21wb25lbnRzLw&ntb=1).

38. What are some other dimension reduction techniques besides PCA?

Ans: other dimension reduction techniques besides PCA: t-SNE( t-distributed Stochastic Neighbor Embedding),linear discriminant analysis (LDA),  singular value decomposition (SVD)

39. Give an example scenario where dimension reduction can be applied.

Ans: **e-mail classification problem**, where we need to classify whether the e-mail is spam or not. This can involve a large number of features, such as whether or not the e-mail has a generic title, the content of the e-mail, whether the e-mail uses a template, etc

Feature Selection:

40. What is feature selection in machine learning?

Ans: Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.

41. Explain the difference between filter, wrapper, and embedded methods of feature selection.

Ans: Filter methods perform the feature selection independently of construction of the classification model. Wrapper methods iteratively select or eliminate a set of features using the prediction accuracy of the classification model. In embedded methods the feature selection is an integral part of the classification model.

42. How does correlation-based feature selection work?

Ans: [Correlation-based Feature Selection (CFS) is a feature selection algorithm that evaluates subsets of features based on the hypothesis that "good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other"](https://www.bing.com/ck/a?!&&p=a03d9632b9997eb4JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY5Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+correlation-based+feature+selection+work%3f&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvRmVhdHVyZV9zZWxlY3Rpb24&ntb=1). [The algorithm is used to reduce dimensionality, remove irrelevant data, and increase learning acc](https://www.bing.com/ck/a?!&&p=040e17c4fe0493ecJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY5OQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=.+How+does+correlation-based+feature+selection+work%3f&u=a1aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2RvY3VtZW50Lzg3MjM5ODA&ntb=1)uracy.

43. How do you handle multicollinearity in feature selection?

Ans: steps to handle multicollinearity in feature selection:

* Understand the extent of multicollinearity Start by calculating correlation
* coefficients or variance inflation factors (VIF) to identify highly correlated features. ...
* Assess feature importance
* Remove redundant features
* Feature transformation
* Regularization techniques

44. What are some common feature selection metrics?

Ans: Feature selection metrics:

Filter method :Information Gain and Chi-square, Pearson’s Correlation Coefficient, Variance Threshold,

Wrapper method: Forward selection, Backward elimination

Embedded Method: Regularization,Tree-based approrach

45. Give an example scenario where feature selection can be applied.

Ans:e.g [SNPs](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism) ,Computer Vision where feature selection can be applied

Data Drift Detection:

46. What is data drift in machine learning?

Ans: [Data drift is a phenomenon in machine learning where the statistical properties of the data used to train a model change over time](https://www.bing.com/ck/a?!&&p=04a369c770f43648JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY5OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+data+drift+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cuZGF0YWNhbXAuY29tL3R1dG9yaWFsL3VuZGVyc3RhbmRpbmctZGF0YS1kcmlmdC1tb2RlbC1kcmlmdA&ntb=1). [This can lead to a decrease in model accuracy over time](https://www.bing.com/ck/a?!&&p=8880554a0eb03fffJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+data+drift+in+machine+learning%3f&u=a1aHR0cHM6Ly9sZWFybi5taWNyb3NvZnQuY29tL2VuLXVzL2F6dXJlL21hY2hpbmUtbGVhcm5pbmcvdjEvaG93LXRvLW1vbml0b3ItZGF0YXNldHM_dmlldz1henVyZW1sLWFwaS0x&ntb=1). [Machine learning models are trained with historical data, but once they are used in the real world, they may become outdated and lose their accuracy over time due to data drift](https://www.bing.com/ck/a?!&&p=18010bdcf439c0e0JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTcwMw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+data+drift+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cuZGF0YWNhbXAuY29tL3R1dG9yaWFsL3VuZGVyc3RhbmRpbmctZGF0YS1kcmlmdC1tb2RlbC1kcmlmdA&ntb=1).

47. Why is data drift detection important?

Ans: Data drift is one of the top reasons model accuracy degrades over time. For machine learning models, data drift is the change in model input data that leads to model performance degradation. Monitoring data drift helps detect these model performance issues.

48. Explain the difference between concept drift and feature drift.

Ans: Feature drift – a change in the probability of p (X), meaning there was a change in the distribution of the model’s input.

X=input ,Y=output

Concept shift/drift happens when posterior probabilities of X and Y, that is the probability of Y as output given X as input changes.

“Data drift”: Covariates P(X)

“Label drift”: Prior probabilities P(Y)

49. What are some techniques used for detecting data drift?

Ans: Some techniques used for detecting data drift

* Population Stability Index: The PSI is a measure of the change in the distribution of a feature between the training and test data
* Kolmogorov-Smirnov: The Kolmogorov-Smirnov test is a non-parametric test that can be used to determine whether two samples come from the same distribution.
* Page-Hinkley method: is a statistical method used to detect changes in the mean of a series of data over time.

50. How can you handle data drift in a machine learning model?

## Ans: Steps:

## Check the data quality

## Investigate the drift

## **You might look at the source of drift and decide to live with it.** Retrain it,

## Calibrate or rebuild the mode: **To rebuild or calibrate the model, you can make more changes to the training pipeline.**

* You might also test some more specific ideas to combat drift:
* Reweigh samples in the training data, giving more importance to the recent ones. The goal is to make the model give priority to newer patterns.
* Identify new segments where the model fails, and create a different model for it. Consider using an ensemble of several models for different segments of the data.
* Change the prediction target. For example, switch from weekly to daily forecast or replace the regression model with classification into categories from "high" to "low."
* Pick a different model architecture to account for ongoing drift: consider incremental or online learning, where the model continuously adapts to new data.
* Apply domain adaptation strategies: There is a number of approaches to help the model better generalize to a new target domain.

## Pause the model and use a fallback

## Find the low-performing segments

## Apply business logic on top of the model**: Setting a new**[**decision threshold**](https://www.evidentlyai.com/classification-metrics/classification-threshold)**is possible for classification problems.**

Data Leakage:

51. What is data leakage in machine learning?

Ans: [Data leakage in machine learning is when information from outside the training data is used to create the model](https://www.bing.com/ck/a?!&&p=fd3bcd65e84b7297JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc4NA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+data+leakage+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cubGlub2RlLmNvbS9kb2NzL2d1aWRlcy93aGF0LWlzLWRhdGEtbGVha2FnZS8&ntb=1). This can lead to overfitting and poor generalization. [Some examples of data leakage are](https://www.bing.com/ck/a?!&&p=f128ef9ddc856a12JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc4Ng&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+data+leakage+in+machine+learning%3f&u=a1aHR0cHM6Ly9tbGlucHJvZHVjdGlvbi5jb20vZGF0YS1sZWFrYWdlLw&ntb=1):

* [Using input data and target that are related in a trivial way, such as historical log data for student attrition](https://www.bing.com/ck/a?!&&p=9d069f260d40d54fJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc5MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+data+leakage+in+machine+learning%3f&u=a1aHR0cHM6Ly9tbGlucHJvZHVjdGlvbi5jb20vZGF0YS1sZWFrYWdlLw&ntb=1)
* [Sharing data from the test dataset with the training dataset, causing the model to perform better than it should](https://www.bing.com/ck/a?!&&p=418ec16b4a8df72cJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc5Mw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+data+leakage+in+machine+learning%3f&u=a1aHR0cHM6Ly93d3cubGlub2RlLmNvbS9kb2NzL2d1aWRlcy93aGF0LWlzLWRhdGEtbGVha2FnZS8&ntb=1)
* [Splitting a time-series dataset randomly instead of chronologically, causing future data to leak into the past](https://www.bing.com/ck/a?!&&p=00ec0244cc31673bJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTc5NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is+data+leakage+in+machine+learning%3f&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvTGVha2FnZV8obWFjaGluZV9sZWFybmluZyk&ntb=1) ID number

52. Why is data leakage a concern?

Ans: [Data leakage in machine learning is the use of **information in the model training process which would not be expected to be available at prediction time**](https://www.bing.com/ck/a?!&&p=4365845e779db458JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1NQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Why+is+data+leakage+a+concern+in+ML%3f&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvTGVha2FnZV8obWFjaGluZV9sZWFybmluZyk&ntb=1). [This causes the predictive scores to overestimate the model's utility when run in a production environment](https://www.bing.com/ck/a?!&&p=e0ee766bce820a34JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1Nw&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Why+is+data+leakage+a+concern+in+ML%3f&u=a1aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvTGVha2FnZV8obWFjaGluZV9sZWFybmluZyk&ntb=1). [Data leakage is one of the major problems in machine learning and occurs when the data used to train an ML algorithm has the information the model is trying to predict](https://www.bing.com/ck/a?!&&p=5acc8bcc660fdb67JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY1OQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Why+is+data+leakage+a+concern+in+ML%3f&u=a1aHR0cHM6Ly93d3cuamF2YXRwb2ludC5jb20vZGF0YS1sZWFrYWdlLWluLW1hY2hpbmUtbGVhcm5pbmc&ntb=1). [It can cause unpredictable and bad prediction outcomes after model deployment](https://www.bing.com/ck/a?!&&p=7607538bab4a4760JmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTY2MQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Why+is+data+leakage+a+concern+in+ML%3f&u=a1aHR0cHM6Ly93d3cuamF2YXRwb2ludC5jb20vZGF0YS1sZWFrYWdlLWluLW1hY2hpbmUtbGVhcm5pbmc&ntb=1).

53. Explain the difference between target leakage and train-test contamination.

Ans: A target leak occurs when your predictors include data that will not be available at the time you make the predictions. Data leakage generally occurs when our training data is fed with the information about the target, but similar data is available when the model is usedin predictions. train-test contamination occurs when you are not careful to distinguish training data from validation data. Validation is meant to be a measure of how well the model performs on data it has not previously considered. You can subtly corrupt this process if the validation data affects preprocessing behaviour. This is referred to as train-test contamination.

54. How can you identify and prevent data leakage in a machine learning pipeline?

Ans:

* Normalizing correctly before cross-validation It is easy to leak data when preparing data for machine learning. As a result, overfitting occurs in training data
* Splitting dataset As mentioned previously, it is normal to split a dataset into training and test sets
* Eliminating duplicates Predictive models always go through a form of data preparation.

55. What are some common sources of data leakage?

Ans: Some common sources of data leakage are:

* Leakage of data from the test set to the training set
* Reversing obfuscation, randomisation or anonymisation of data that was intentionally included
* Inclusion of information from data samples outside algorithm’s scope for the intended use
* Inclusion of data not present in the model’s operational environment.

56. Give an example scenario where data leakage can occur.

Ans: Scenario where data leakage can occur:

* Sending information to the wrong recipient Ouch, simple yet effective and the biggest cause of data leaks: sending sensitive data to the wrong person. ...
* Email addresses of all recipients in CC Whoever puts all addresses in the cc, makes all recipients in that group public
* Unsafe servers
* Weak passwords
* Lack of the right encryption

Cross Validation:

57. What is cross-validation in machine learning?

Ans: Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times, each time using a different fold as the validation set. Finally, the results from each validation step are averaged to produce a more robust estimate of the model’s performance.

58. Why is cross-validation important?

Ans: Cross-validation is important to prevent overfitting, which occurs when a model is trained too well on the training data and performs poorly on new, unseen data. By evaluating the model on multiple validation sets, cross validation provides a more realistic estimate of the model’s generalization performance, i.e., its ability to perform well on new, unseen data.

59. Explain the difference between k-fold cross-validation and stratified k-fold cross-validation.

Ans: [KFold is a cross-validator that divides the dataset into k folds. Stratified KFold is a variation of KFold that ensures that each fold of dataset has the same proportion of observations with a given label](https://www.bing.com/ck/a?!&&p=19665eeb796d3fcdJmltdHM9MTY4OTEyMDAwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTczOQ&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=Explain+the+difference+between+k-fold+cross-validation+and+stratified+k-fold+cross-validation&u=a1aHR0cHM6Ly9zdGFja292ZXJmbG93LmNvbS9xdWVzdGlvbnMvNjUzMTg5MzEvc3RyYXRpZmllZGtmb2xkLXZzLWtmb2xkLWluLXNjaWtpdC1sZWFybg&ntb=1).

60. How do you interpret the cross-validation results?

Ans: Cross Validation results can be interpreted in following ways as well as with statisctics:

* Continuous Ranked Probability Score—A positive number measuring the accuracy and precision of the predicted value, where a smaller value is better.
* Validation Quantile—The quantile of the measured value with respect to the predictive distribution. If the model is correctly configured, the validation quantiles will be uniformly distributed between 0 and 1 and show no patterns.
* Inside 90 Percent Interval—An indicator (1 or 0) of whether the measured value is within a 90 percent prediction interval (analogous to a confidence interval). If the model is correctly configured, approximately 90 percent of points will fall inside the interval and have the value 1.
* Inside 95 Percent Interval—An indicator (1 or 0) of whether the measured value is within a 95 percent prediction interval. If the model is correctly configured, approximately 95 percent of points will fall inside the interval and have the value 1.
* In cross validation ,we are having Mean,Standard Error(the error divided by the standard error), predicted, Measured, Standarized Error(he error divided by the standard error.)