
PREPARE FOR
DATA SCIENCE
INTERVIEW THE
RIGHT WAY
205+ MCQ TO TEST
YOUR KNOWLEDGE
EDCORNER LEARNING

Prepare For Data Science Interview the Right Way

Prepare For Data Science Interview the Right Way

Edcorner Learning

Table of Contents

Introduction
Module 1: Interview Preparation Part 1
Module 2: Interview Preparation Part 2
Module 3: Interview Preparation Part 3
Module 4: Interview Preparation Part 4
Module 6: Interview Preparation Part 5

Introduction

what is data science?

Data science is a field of computer science (in a general sense) that applies statistical tools to study data. Data science encompasses many fields such as statistics, economics, operations research, statistics and probability, machine learning, computer vision, and pattern recognition.

A data scientist isn't necessarily a statistician but may use statistics to improve their skills. The mathematical theories of statistics are often used as tools for data science, however, statistical tools are often applied through other fields such as computer science.

Data science as a field has only emerged over the past decade or so and has become very popular due to the large amounts of data that have been collected by modern technology.

There is no shortage of data in the world; we often wonder what could be done with it, or if it will be useful at all. The only shortage we seem to have is the time to truly study it.

This Book consist of 205+ MCQ's of data science concepts with answers to maintain difficulty level and capture all areas in best way as possible.

Please Note answer to each question is on next page of the questions, First try to think answer by yourself and check to confirm your answers.

This book is mainly designed to check your understanding of data scientist concepts and its depth before you appear in the interview, all in all you can improve your performance and knowledge the right way.

Module 1: Interview Preparation Part 1

Question #:1 What is feature selection?

1. In machine learning, feature selection is the process of selecting a subset of appropriate features (variables) to be used in machine learning model.

2. In machine learning, feature selection is the process of selecting the appropriate hyperparameters for a machine learning model.

Question #:2 What is the gradient descent algorithm?

1. The gradient descent algorithm is a popular technique for selecting optimal variables for a machine learning model.

2. Gradient descent is a popular optimization technique in machine learning and deep learning that can be used with most machine learning algorithms (helps train the model).

3. The gradient descent algorithm is a popular technique in machine learning for dimensionality reduction.

Question 1	Answer - 1
Question 2	Answer - 2

Question #:3 We have the following sentence:
'Python is a programming language'
Select all bigrames from the given sentence.

1. 'Python is', 'a programming', 'language'
2. 'Python is', 'is a', 'a programming', 'programming language'
3. 'Python', is', 'a', 'programming', 'language'
4. 'Python is a programming language'

Question #:4 What is the goal of ensemble methods?

1. The goal of ensemble methods is to combine the predictions of several underlying estimators built with a given machine learning algorithm to improve generalization/robustness to a single estimator.
2. The goal of ensemble methods is to break down data into multiple groups so that points in the same groups are more similar to other points in the same group than in other groups.

Question 3	Answer - 2
Question 4	Answer - 1

Question #:5 What is data anonymization?

1. Data anonymization consists of splitting the data into a training set and a testing set.

2. Data anonymization is a type of data processing that aims to protect privacy. It is the process of removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous.

3. Data anonymization is a type of data processing that aims to standardize/normalize numerical variables.

Question #:6 Select algorithms that can be used in binary classification problems.

1. Logistic regression
2. Neural networks
3. Linear regression
4. Random forests
5. Decision trees

Question 5	Answer - 2
Question 6	Answer - 1,2,4,5

Question #:7 What does NLP stand for in machine learning?

1. Network Layer Protocol

2. Natural Language Processing. It is a branch of artificial intelligence that deals with natural language processing.

3. Nonlinear Programming

4. Neuro-Linguistic Programming

Question #:8 Select true statements about decision tree algorithms.

1. In classification problems, the goal of the decision tree algorithm is to create a model that predicts the value of the target variable for which the decision tree uses the tree representation to solve the problem where the leaf node corresponds to the class label and the attributes are represented on the inner node of the tree.

2. Decision tree algorithms belong to the family of unsupervised machine learning algorithms.

3. Decision tree algorithms belong to the family of supervised machine learning algorithms

4. Decision tree algorithms can be used both for a classification problem and for a regression problem.

Question 7	Answer - 2
Question 8	Answer - 1,3,4

Question #:9 Why is the Naive Bayes classifier called 'naive'?

1. Since it is based on Bayes' theorem with the assumption of independence of predictors, hence the term 'naive'. Put simply, the classifier assumes that the presence of a certain feature in the class is not related to the presence of any other feature.
2. The algorithm was named after its creator - Naive-Bayes.

Question #:10 The following random variable X with a discrete distribution is given:
P(X = 1) = 0.2

1. P(X = 3) = 0.4
2. P(X = 6) = 0.4
3. The expected value of a random variable X is:
4. 6.0
5. 3.8
6. 3.0
7. 3.5

Question 9	Answer - 1
Question 10	Answer - 5

Question #:11 What is multiclass classification?

1. In machine learning, multiclass classification is a classification task that assigns to each sample multiple labels. For example, we can assign two labels to a photo: item and color (dress, green).

2. In machine learning, multiclass classification divides the elements of a set into two groups based on a classification rule.

3. In machine learning, multiclass classification is the problem of classifying into one of three or more classes.

Question #:12 What is semi-supervised learning?

1. Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data while learning. Semi-supervised learning falls between unsupervised and supervised learning.

2. Semi-supervised learning is a machine learning task that trains a function that maps input to output based on labeled training data.

3. Semi-supervised learning is a type of machine learning whose task is to discover patterns in a data set without pre-existing labels.

Question 11	Answer - 3
Question 12	Answer - 1

Question #:13 Explain the basic principle of the nearest neighbors algorithms.

1. The main purpose of the algorithms in this group is to find a predetermined number of centroids and assign the nearest points to them based on distance or local density.

2. The main purpose of this group of algorithms is to find a predetermined number of training samples closest to a new point and use them to predict the label. The number of samples can be a user-defined constant or vary based on local point density. The distance can be any metric, the standard Euclidean distance is most often chosen.

Question #:14 What is an outlier?

1. An outlier is an observation that lies a normal distance from other values in a random sample from a population.

2. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Question 13	Answer - 2
Question 14	Answer - 2

Question #:15 Select advantages of decision trees.

1. Simple to understand and interpret.
2. White box model -> it is possible to explain the model result.
3. Simple to visualize.
4. Black box model -> it is impossible to explain the model result.
5. Requires little data preparation.

Question #:16 Select feature selection methods:

1. Grid Search
2. Filter methods (ANOVA, Pearson correlation)
3. Embedded Methods
4. Wrapper methods (forward selection, backward selection, stepwise selection)

Question 15	Answer - 1,2,3,5
Question 16	Answer - 2,3,4

Question #:17 Select the mode of the following data set:
3, 5, 2, 6, 7, 3, 0, 3
3
(Correct)
7
6
3.625

Question #:18 What is a time series?

1. A time series is a sequence of data points that occur in successive order over some period of time. Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time.

2. A time series is a sequence of data that is not ordered over time.

Question 17	Answer - 1
Question 18	Answer - 1

Question #:19 What is multi-label classification?

1. In machine learning, multi-label classification is the problem of classifying into one of three or more classes.

2. In machine learning, multi-label classification is a classification task that assigns to each sample multiple labels. For example, we can assign two labels to a photo: item and color (dress, green).

Question #:20 What is EDA - Exploratory Data Analysis?

1. Exploratory data analysis is the entire process of preparing data for a machine learning model (filling missing values, standardization, normalization, feature engineering, splitting into training/validation/testing set, etc.).

2. In statistics, exploratory data analysis is the approach of analyzing data to summarize main characteristics, often using data visualization. The EDA is used to test what the data can tell us beyond formal modeling or hypothesis testing.

Question 19	Answer - 2
Question 20	Answer - 2

Question #:21 Select the mode of the following data set:
3, 5, 2, 6, 7, 3, 0, 6
4
3
6
The mode of the given dataset cannot be clearly identified.

1. 3 and 6

Question #:22 Select median of the following dataset:
5, 2, 6, 8
5.25

1. 5.5
2. 5
3. 6

Question 21	Answer - 1
Question 22	Answer - 1

Question #:23 Select median of the following dataset:
5, 2, 6, 8, 9, 1, 18

1. 6
2. 7
3. 7.5
4. 5

Question #:24 What is PCA - Principal Component Analysis?

1. Principal Component Analysis is the approach of analyzing data sets to summarize their main characteristics, often using data visualization. It is used to test what the data can tell us beyond formal modeling or hypothesis testing.

2. Principal Component Analysis is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Question 23	Answer - 1
Question 24	Answer - 2

Question #:25 Select measures of central tendency that are insensitive to outliers.

1. Median
2. Variance
3. Arithmetic mean
4. Standard deviation
5. Covariance

Question #:26 You work in an e-commerce company and have to split clients into groups of similar clients. Select the clustering algorithms you can consider when building the model.

1. K-Means algorithm
2. Agglomerative Clustering
3. DBSCAN
4. Random forest
5. Mean-Shift Clustering
6. Decision tree

Question 25	Answer - 1
Question 26	Answer - 1,2,3,5

Question #:27 Select true statements regarding SGD (Stochastic Gradient Descent).

1. The SGD algorithm requires many hyperparameters.
2. The SGD algorithm is easy to implement.
3. The SGD algorithm does not require any hyperparameter.
4. SGD algorithm is sensitive to feature scaling.
5. The SGD algorithm is quite effective.

Question #:28 Select reasons for the feature selection.

1. Feature selection extends the training time of the model.
2. Feature selection allows us to reduce dimensionality.
3. Feature selection allows us to train the model shorter.
4. Feature selection allows us to simplify the model in order to facilitate its interpretation.

Question 27	Answer - 1,2,4,5
Question 28	Answer - 2,3,4

Question #:29 What is the random forest algorithm?

1. Random forests are an ensemble method of classification, regression, and other tasks that involves constructing multiple decision trees while learning. In the case of classification tasks, the result of a random forest is the class chosen by most trees. For regression tasks, the average of each tree's prediction is returned.

2. Random forests are the basic unsupervised learning algorithm. It is used to extract association rules.

3. Random forests are the basic unsupervised learning algorithm. It is used to group data together to find similar groups.

Question #:30 Select disadvantages of decision trees.

1. Decision trees cannot be visualized.

2. It is a black-box model. It is impossible to explain the result of the model.

3. Decision trees are easy to overfitting.

4. Decision trees can be unstable because slight differences in data can generate a completely different tree.

Question 29	Answer - 1
Question 30	Answer - 3,4

Question #:31 What is mode?

1. In probability theory and statistics, a mode is a measure of the dispersion of a random variable, which means that it is a measure of how far the values of a random variable are distributed from its mean value.

2. Mode is a measure of the central tendency. It is the value of the feature that divides the ordered statistical population into two equal parts in such a way that there is an equal number of observations below and above this value.

3. Mode is the value that appears most frequently in a set of data values.

Question #:32 What is overfitting?

1. Overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data.

2. Overfitting is a concept in data science that occurs when the data model is unable to accurately capture the relationship between the input and output data, generating a high error rate in both the training set and the test data.

Question 31	Answer - 3
Question 32	Answer - 1

Question #:33 Select true statements about covariance.

1. Covariance is positive if the higher values of one variable mainly correspond to the higher values of the other variable and the same is true for the lower values.

2. Covariance is positive when the higher values of one variable mainly correspond to the lesser values of the other.

3. Covariance is negative if the higher values of one variable mainly correspond to the higher values of the other variable and the same is true for the lower values.

4. Covariance is negative when the higher values of one variable mainly correspond to the lesser values of the other.

5. In probability theory and statistics, covariance is a measure of the combined variability of two random variables.

Question #:34 What is variance?

1. In probability theory and statistics, variance is a measure of the dispersion of a random variable, which means that it is a measure of how far the values of a random variable are distributed from its mean value.

2. Variance is a measure of central tendency. It is the value of the feature that divides the ordered statistical population into two equal parts in such a way that there is an equal number of observations below and above this value.

3. Variance is the value that appears most frequently in a set of data values.

Question 33	Answer - 1,4,5
Question 34	Answer - 1

Question #:35 What is binary classification?

1. In machine learning, binary classification is the problem of classifying into one of three or more classes.

2. Binary classification is the task of classifying the elements of a set into two groups based on a classification rule.

3. In machine learning, multiclass classification is a classification task that assigns to each sample multiple labels. For example, we can assign two labels to a photo: item and color (dress, green).

Question #:36 What is the main drawback of MSE - Mean Squared Error metric?

1. The MSE metric is a metric that needs a lot of computing power.

2. The MSE metric is not prone to outliers.

3. The MSE metric is prone to outliers and put too much emphasis on large deviations.

Question 35	Answer - 2
Question 36	Answer - 3

Question #:37 If two random variables X, Y are independent then their covariance is equal to:
1
We cannot compute a covariance based on this information.

1. -1
2. 0.5
3. 0

Question #:38 Select true statements about the Multilayer Perceptron (MLP).

1. Training MLP networks is possible thanks to the backpropagation algorithm.
2. Multilayer perceptron is the most popular type of neural network that consists of at least three layers (input layer, hidden layer, output layer).
3. Multilayer perceptron, unlike a single-layer perceptron, can be used to classify sets that are not linearly separable.

Question 37	Answer - 3
Question 38	Answer - 1,2,3

Question #:39 What is hierarchical clustering?

1. In data science and statistics, hierarchical clustering is a cluster analysis method that aims to build a hierarchy of clusters. Hierarchical clustering strategies generally fall into two types: (agglomerative - bottom-up approach, divisive - top-down approach).

2. Hierarchical clustering can be defined as a basic unsupervised learning algorithm that includes defining a predetermined number of centroids to which points are assigned based on a specific similarity.

Question #:40 Select true statements about methods based on Support Vector Machines (SVMs).

1. SVMs are quite versatile (different kernel functions can be specified for the decision function).

2. SVMs are effective in multidimensional spaces.

3. SVMs remain effective in cases where the number of dimensions is greater than the number of samples.

4. SVMs use a subset of training data in a decision function (called support vectors), making it memory-efficient.

5. SVMs do not provide probabilities directly (it can be computed).

Question 39	Answer - 1
Question 40	Answer - 1,2,3,4,5

Question #:41 Select true statements about the median.

1. The median is a measure of central tendency.

2. The median divides the ordered statistical population into two equal parts in such a way that there is an equal number of observations below and above this value.

3. The median is the third quartile of the probability distribution.

4. Median or median value is a measure of the dispersion of a random variable.

5. The median is the second quartile of the probability distribution.

Question 41	Answer - 1,2,5

Module 2: Interview Preparation Part 2

Question #:1 Select true statements about the drawbacks of linear models.

1. Linear models are prone to multicollinearity.
2. Linear models are prone to outliers.
3. Linear models are prone to overfitting.
4. One of the major drawbacks of linear models is the nonlinearity assumption.
5. Linear models are not prone to outliers.
6. One of the major drawbacks of linear models is the linearity assumption.

Question #:2 What is n-gram?

1. In machine learning, an n-gram is the representation of a decision tree model as a graph.
2. In computational linguistics, a n-gram is a continuous sequence of n elements from a given text or speech sample. Elements can be syllables, letters, or words, for example.
3. In machine learning, an n-gram is the representation of a random forest model as a graph.

Question 1	Answer - 1,2,3,6
Question 2	Answer - 2

Question #:3 Select true statements about eigenvectors.

1. Eigenvectors are linearly independent.

2. An eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it.

3. Eigenvectors are linearly dependent.

4. Eigenvectors are used to make linear transformation understandable.

Question #:4 You have to prepare a dataset for a model with 20% of missing data. Select answers that may help you solve this problem.

1. For smaller sets, the missing values can be replaced with e.g. the mean value, median or mode.

2. You can train an additional machine learning model to fill in the missing data.

3. If the set is quite large, you can remove the records with missing data.

4. Missing values can be replaced with e.g. mean value, median or mode calculated on the basis of other known variables (e.g. categorical variables).

Question 3	Answer - 1,2,4
Question 4	Answer - 1,2,3,4

Question #:5 What is the activation function for?

1. The activation function is only used in the last layer of the neural network to calculate the final results of the model.

2. The activation function is used to introduce nonlinearities into the neural network, helping it learn a more complex function. Without the activation function, the neural network would only be able to learn a linear function, which is a linear combination of its inputs.

3. The activation function is used to introduce linearity into the neural network.

Question #:6 Select true statements about the given technologies (Python, R, SAS).

1. The SAS language is free (open source). Mostly used by the academic community. It is the main language for statistical computing.

2. Python is a powerful open source programming language that is easy to learn and works well with most other tools and technologies. Python has many libraries and modules developed by a large community.

3. The R language is free (open source). Mostly used by the academic community. It is the main language for statistical computing.

4. SAS is one of the most frequently used analytical tools used by the largest companies with a fairly long history (banks, insurance). SAS is not free.

Question 5	Answer - 2
Question 6	Answer - 2,3,4

Question #:7 What is univariate analysis?

1. Univariate analysis is one of the simpler forms of statistical analysis and involves two variables to determine the relationship between them.

2. Univariate analysis can help determine how much easier it is to know and predict the value of one variable if you know the value of the other variable.

3. As the name suggests, this is a single variable analysis methodology. The simplest case in statistical analysis.

4. Univariate analysis involves many statistical methods that are designed to take into account many variables and examine the contribution of each.

Question #:8 What do you mean by normal distribution (Gaussian distribution)?

1. The normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. The density of the normal distribution is represented as a bell curve.

2. The normal distribution is a discrete probability distribution that describes the number of k successes over n independent trials, each of which has a constant success probability of p.

3. It is a continuous probability distribution for which the probability density in a given interval is constant and not equal to zero.

Question 7	Answer - 3
Question 8	Answer - 1

Question #:9 What is collaborative filtering?

1. Collaborative filtering is a technique used to select features for a machine learning model.

2. Collaborative filtering is a popular clustering algorithm.

3. Collaborative filtering is a technique used by recommendation systems. It is a method of making automatic predictions about the interests of a user by collecting preferences or taste information from many users.

Question #:10 What is reinforcement learning (RL)?

1. Reinforcement learning is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize cumulative reward.

2. Reinforcement learning is a type of machine learning that aims to find a function that maps inputs to outputs based on the labeled data.

3. Reinforcement learning is a type of machine learning that aims to discover new patterns in a data set.

Question 9	Answer - 3
Question 10	Answer - 1

Question #:11 What is the K-means clustering algorithm?

1. K-Means algorithm is an ensemble learning method for classification, regression, and other tasks that involves constructing multiple decision trees while learning.

2. K-Means algorithm is one of the statistical techniques used to predict a binary result.

3. K-means algorithm is an iterative algorithm that tries to partition the dataset into K distinct non-overlapping clusters where each data point belongs to only one group.

Question #:12 What is multivariate analysis?

1. Multivariate analysis is one of the simpler forms of statistical analysis and involves analyzing exactly two variables to determine the relationship between them.

2. Multivariate analysis includes many statistical methods that are designed to allow us to include multiple variables and examine the contribution of each.

3. Multivariate analysis is an analysis methodology based on the analysis of one variable. This is the simplest case in statistical analysis.

Question 11	Answer - 3
Question 12	Answer - 2

Question #:13 How can we avoid overfitting?

1. Keep the model simple - consider fewer variables, thus removing some of the noise in the data.
2. Use data augmentation techniques.
3. Use a cross-validation technique.
4. Use regularization techniques.

Question #:14 Standard deviation can be negative [True/False].

1. True
2. False

Question #:15 Select true statements about A/B testing.

1. A/B testing is a way to compare two versions of one variable, usually by testing a user's response to variant A with variant B and determining which of the two options is more effective.
2. A/B testing consists of a randomized experiment with two variants, A and B.
3. A/B testing is a methodology for testing user experience.

Question 13	Answer - 1,2,3,4
Question 14	Answer - 2
Question 15	Answer - 1,2,3

Question #:16 Select true statements about the difference between supervised learning and unsupervised learning.

1. Linear regression is an example of supervised learning.

2. The K-Means clustering algorithm is an example of supervised learning.

3. Logistic regression is an example of unsupervised learning.

4. The K-Means clustering algorithm is an example of unsupervised learning.

5. Unsupervised learning is a type of machine learning in which the algorithm is not provided with any pre-assigned labels or scores for the training data.

6. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

Question #:17 Select popular tools for working with neural networks.

1. Flask

2. Django

3. Pytorch

4. Keras

5. Tensorflow

Question 16	Answer - 1,4,5,6

Question 17	Answer - 3,4,5

Question #:18 Select methods used to select the correct features for the model.

1. Chi-Square
2. LDA (Linear Discriminant Analysis)
3. ANOVA

Question #:19 What is dimensionality reduction?

1. Dimensionality reduction is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful information of the original data.

2. Dimensionality reduction is the transformation of data from a low-dimensional space into a high-dimensional space so that the high-dimensional representation retains some meaningful information of the original data.

Question 18	Answer - 1,2,3
Question 19	Answer - 1

Question #:20 Select popular techniques for dimensionality reduction.

1. t-SNE (T-distributed Stochastic Neighbor Embedding)
2. LDA (Linear Discriminant Analysis)
3. Backward Elimination
4. Logistic regression
5. FA (Factor Analysis)
6. PCA (Principal Component Analysis)

Question #:21 What is interpolation?

1. Interpolation is the process of estimating unknown values that fall between known values.

2. Interpolation is a type of estimation that estimates values outside the original observation range.

Answer: 1

Question 20	Answer - 1,2,3,5,6
Question 21	Answer - 1

Question #:22 What is the Euclidean distance of the given points?

1. A = (0, 3)
2. B = (4, 0)
3. 3
4. The Euclidean distance cannot be calculated.
5. 0
6. 4
7. 5

Question #:23 What is clustering?

1. Clustering is the task of dividing data into multiple groups so that items in the same groups are more similar to other items in the same group than in other groups.

2. Clustering is the task of dividing the elements of a set into two groups based on a classification rule.

3. Clustering is the task of extracting association rules.

Question 22	Answer - 7
Question 23	Answer - 1

Question #:24 What do you mean by linear regression?

1. The linear regression method uses a tree representation to solve a problem.

2. The linear regression is an ensemble learning method that involves building multiple decision trees while learning.

3. It is the most widely used method of predictive analysis. The linear regression method is used to describe the relationship between a dependent variable and one or more independent variables. Graphically, a linear regression model in (R^2) can be represented as a single line fit on a scatter plot.

4. The linear regression is one of the statistical techniques for predicting a binary score. An example is the verification of the creditworthiness of a bank customer (default: yes/no).

Question #:25 We need to build a model that determines the probability if the client will resign from telecom services (churn). Which of the following algorithms can we consider when creating the model?

1. K-means clustering

2. Logistic regression

3. Decision tree

4. Linear Regression

5. PCA algorithm

6. Apriori algorithm

Question 24	Answer - 3

Question 25	Answer - 2,3

Question #:26 Do gradient descent algorithms converge to the same point all the time?

1. Not always. Gradient descent converges to a local minimum if it starts close enough to that minimum. If there are multiple local minimums, its convergence depends on where the iteration starts.

2. Yes.

Question #:27 Which type of estimation is more risky, interpolation or extrapolation?

1. interpolation

2. extrapolation

Question #:28 What is extrapolation?

1. Extrapolation is the process of estimating unknown values that fall between known values.

2. Extrapolation is a type of estimation that estimates values outside the original observation range.

Question 26	Answer - 1
Question 27	Answer - 2
Question 28	Answer - 2

Question #:29 What are recommendation systems?

1. Recommendation systems are currently widely used in many areas, such as movie recommendations (Netflix), music (Spotify), tags (Twitter), search queries (Google) or products (Amazon). Recommendation systems work based on a user's past behavior in order to build a model for the future. This allows you to predict future product purchases, film or music preferences.

2. Recommendation system is a software for tracking changes in any set of files, usually used for coordinating work among programmers collaboratively developing source code during software development

3. Recommendation systems are systems that automatically adjust the appropriate algorithm to any machine learning problem (recommendation model).

Question #:30 What do you mean by feature vector?

1. In machine learning, a feature vector is an n-dimensional vector that stores the names of features used in the machine learning model.

2. In pattern recognition and machine learning, a feature vector is an n-dimensional vector of numerical features that represent some object. Many algorithms in machine learning require a numerical representation of objects, since such representations facilitate processing and statistical analysis.

3. In neural networks, a feature vector is a vector of intercepts.

Question 29	Answer - 1
Question 30	Answer - 2

Question #:31 What is machine learning?

1. Machine learning systems are systems that operate on the basis of the developer's expertise (knowledge is not extracted from the data, but implemented manually by the developer).

2. Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to predict the future.

Question #:32 What is deep learning?

1. Deep learning systems are systems that operate on the basis of the developer's expertise (knowledge is not extracted from the data, but implemented manually by the developer).

2. Deep learning is the term used to describe machine learning models that are not built using neural networks.

3. Deep learning is part of a wider family of machine learning methods based on artificial neural networks. Deep learning can be supervised, semi-supervised or unsupervised.

Question #:33 What is bivariate analysis?

1. Bivariate analysis is an analysis methodology based on the analysis of one variable. This is the simplest case in statistical analysis.

2. Bivariate analysis is one of the simpler forms of statistical analysis and involves the analysis of two variables to determine the relationship between them. Bivariate analysis can help determine how much easier it is to predict the value of one variable if we know the value of the other variable.

Question 31	Answer - 2
Question 32	Answer - 3
Question 33	Answer - 2

Question #:34 Select programming languages used in data science that are open source.

1. Python
2. SAS
3. R

Question #:35 Select the answer containing only bigrames.

1. 'she likes', 'headphone set', 'video games', 'stock market'
2. 'she likes IT', 'headphone set was', 'my video games', 'stock market cap'
3. 'she', 'likes', 'headphone', 'set', 'video', 'games', 'stock', 'market'
4. 'she likes', 'she', 'likes', 'headphone set', 'headphone', 'set'

Question #:36 What is the Internet of Things (IoT)?

1. The Internet of Thing describes a network of physical objects - things - that are embedded in sensors, software, and other technologies to connect and exchange data with other devices and systems over the Internet.
2. The Internet of Things is a technology that superimposes a computer generated image on a user's view of the real world, thus providing a composite view.

Question 34	Answer - 1,3
Question 35	Answer - 1
Question 36	Answer - 1

Question #:37 What is the purpose of cross-validation?

1. The purpose of cross-validation is to test many different values for the hyperparameters of the machine learning algorithm.

2. The purpose of cross-validation is to test the ability of a machine learning model to predict on new, unseen data. It is also used to detect problems such as overfitting and underfitting.

3. The purpose of cross-validation is to extract relevant variables for the machine learning model.

Question #:38 What is logistic regression for?

1. The logistic regression is one of the statistical techniques for predicting a binary result. An example is the verification of the creditworthiness of a bank customer (default: yes/no).

2. The logistic regression is an ensemble learning method that involves creating multiple decision trees while learning.

3. The logistic regression method uses a tree representation to solve a problem.

4. The logistic regression method is used to describe the relationship between a dependent variable and one or more independent variables. Graphically, a logistic regression model can be represented as a single line fit on a scatter plot.

Question 37	Answer - 2
Question 38	Answer - 1

Question #:39 Select true statements about data cleansing.

1. Proper data cleansing can significantly affect the quality of the machine learning model.

2. In data science, it is important to ensure that the data is good enough (clean) for analysis.

3. Proper data cleansing cannot significantly affect the quality of the machine learning model.

4. Data cleansing requires a tremendous amount of time and effort due to the multiple sources from which the data can derive.

5. Data cleansing is an essential part of data science as data can be prone to error due to human neglect, corruption in transmission, or many other things.

Question #:40 What are RNN (Recurrent Neural Network) networks for?

1. A recurrent neural network is a type of artificial neural network that consists of one or more convolutional layers.

2. A recurrent neural network is a type of artificial neural network that contains one or more pooling layers.

3. A recurrent neural network is a type of artificial neural network which uses sequential data or time series data. In this type of neural network, the output from the current step is given as input to the next step.

Question 39	Answer - 1,2,4,5
Question 40	Answer - 3

Question #:41 You work in an e-commerce company and need to group customers with similar behavior in order to optimize your advertising and mailing campaigns. Which of the following algorithms can you consider when creating the model?

1. DBSCAN
2. Linear regression
3. K-means clustering
4. Decision tree
5. Random forest
6. PCA algorithm

Question 41	Answer - 1,3

Module 3: Interview Preparation Part 3

Question #:1 What is a batch in machine learning terminology?

1. The dataset used in one iteration (i.e. one gradient update) of model training.

2. Batch is one pass of all training data across the network. For example, a batch set to 20 means the training data runs 20 times through the neural network during training.

Question #:2 What are the checkpoints when training the neural network?

1. The checkpoint allows you to save the currently created neural network architecture.

2. Checkpoints capture the exact value of all parameters used by the model. Checkpoints allow us to save and export model weights.

Question #:3 What is feature engineering?

1. The process of determining which features can be useful in model training. Feature engineering is also based on the use of domain knowledge to extract features (characteristics, properties, attributes) from raw data.

2. It is a process consisting of the appropriate selection of hyperparameters for the machine learning model.

Question 1	Answer - 1
Question 2	Answer - 2
Question 3	Answer - 1

Question #:4 Select true statements about the baseline model in machine learning.

1. A baseline model is a model that is used as a benchmark to compare the performance of another model, usually a more complex one.

2. The baseline model helps data scientists quantify the minimum expected performance that a new model must achieve for the new model to be useful.

3. For example, a logistic regression model can serve as a good benchmark - a base model for neural networks.

4. In classification problems, a linear regression model can be a good reference point - the base model for neural networks.

Question #:5 What is the stride in the convolution layer of the CNN network for?

1. Stride determines the size of the convolutional filter. For example, a stride set to 3 means we are using a 3x3 convolution filter.

2. Stride is the offset step size of convolutional filters. It is usually set to 1, but can be 2-6 or even higher to improve computing performance.

3. Stride determines the size of pooling. For example, a stride set to 3 means we are using a 3x3 pooling operation.

Question 4	Answer - 1,2,3
Question 5	Answer - 2

Question #:6 What are categorical features?

1. Categorical features represent numerical values that can be measured and logically ordered.

2. Features having a discrete set of possible values. For example, eye color: blue, green, hazel, brown, etc.

Question #:7 Consider the problem of binary classification. We build a model to predict whether a person suffers from a certain disease or not. What is the type II error for this problem?

1. Type II error: A sick person is incorrectly classified as healthy.

2. Type II error: A healthy person is incorrectly classified as sick.

Question #:8 What is the confirmation bias in machine learning?

1. It is a tendency to critically verify information that contradicts previous opinions, while uncritically accepting information that confirms them.

2. It is a tendency to judge things from the standpoint of your profession, ignoring the larger viewpoint.

3. It is a tendency to search for, interpret, and favor information in a way that confirms previous beliefs or hypotheses. Data Scientists may inadvertently collect or label data in a way that influences the outcome that supports their existing beliefs.

Question 6	Answer - 2
Question 7	Answer - 1
Question 8	Answer - 3

Question #:9 Select true statements about the confusion matrix.

1. A confusion matrix for a multi-class classification problem can help identify error patterns. For example, in the classic handwritten digit recognition model, the confusion matrix may reveal that the trained model tends to mis-predict 1 instead of 7.

2. In machine learning, it is the covariance matrix of the variables used to build the machine learning model.

3. In machine learning, it is a matrix that summarizes the effectiveness of the classification model's predictions. One axis of a confusion matrix is the label that the model predicted, and the other axis is the actual label.

Question #:10 Select true statements about the model's convergence.

1. In deep learning, the loss sometimes remain constant or nearly the same for many iterations before finally descending, which can get a false sense of convergence.

2. The model converges when additional training on the current data does not improve the model.

Question #:11 How does the Maximum Residual Error (MRE) metric work in regression problems?

1. MRE is the metric that returns the maximum residual error. This is the worst case of error between the predicted value and the true value.

2. MRE is a measure of error calculated from the mean of the absolute errors.

3. MRE is a metric that calculates the mean squared error.

Question 9	Answer - 1,3
Question 10	Answer - 1,2
Question 11	Answer - 1

Question #:12 What is Augmented Reality (AR) technology?

1. Technology that categorizes pixels into a specific category, such as a car, road, sign, or pedestrian.

2. Technology that superimposes a computer generated image on a user's view of the real world, thus providing a composite view.

3. Technology that is used to determine the content of an image. For this purpose, neural networks (mainly CNN) can be trained to identify various objects (the classic example of a dog vs. a cat).

Question #:13 What is data augmentation?

1. Increasing the number of examples in the training set by transforming existing data (e.g. image zoom, image rotation, image stretching).

2. It is a technique that generates images through an artificial neural network.

Question 12	Answer - 2
Question 13	Answer - 1

Question #:14 What is variable clipping?

1. One of the techniques for dealing with outliers. It consists of reducing the value of a variable that is greater than the set maximum value to that maximum value and increasing the value of the variable that is less than the specified minimum value to this minimum value.

2. Variable clipping consists of transforming a continuous variable into a discrete variable.

Question 14	Answer - 1

Question #:15 Consider the problem of binary classification. We build a model to predict whether a person suffers from a certain disease or not. What is the type I error for this problem?

1. Type I error: A sick person is incorrectly classified as healthy.
2. Type I error: A healthy person is incorrectly classified as sick.

Question #:16 What does the abbreviation AR stand for in technologies?

1. Additional Reality
2. Augmented Reality
3. Anomaly Report

Question #:17 What is a hidden layer in a neural network?

1. The hidden layer is the first layer of the neural network.
2. The hidden layer is the last layer of the neural network.
3. It is the layer in the neural network between the input layer (features) and the output layer (predictions). Hidden layers usually include an activation function. A deep neural network contains at least one hidden layer.

Question 15	Answer - 2
Question 16	Answer - 2
Question 17	Answer - 3

Question #:18 How does the Mean Squared Error (MSE) metric work in regression problems?

1. The MSE is the metric that returns the maximum residual error. This is the worst case of error between the predicted value and the true value.

2. MSE is a metric that calculates the mean squared error. It measures the average of the squares of the errors (difference between the estimated values and the actual value).

3. MSE is a measure of error calculated from the mean of the absolute errors. MSE is the mean absolute difference between the expected and predicted values in all training examples.

Question #:19 Consider the problem of binary classification. We build a model to predict whether a person suffers from a certain disease or not. What is more dangerous in the consequences of committing a type I error or a type II error?

1. Committing a type I error - we will diagnose a healthy person as sick -> after additional tests, such a case can be eliminated

2. Committing a type II error - we will not correctly diagnose a sick person -> the person may not receive help on time

Question 18	Answer - 2
Question 19	Answer - 2

Question #:20 Select the metrics that can be used for regression problems.

1. Mean Squared Error (MSE)
2. Confusion matrix
3. Maximum Residual Error (MRE)
4. Mean Absolute Error (MAE)
5. Accuracy
6. R^2 Score

Question #:21 Consider the problem of binary classification. We build a model to predict whether a person suffers from a certain disease or not. Which is milder in consequences, committing a type I error or a type II error?

1. Committing a type II error - we will not correctly diagnose a sick person -> may not receive help on time

2. Committing a type I error - we will diagnose a healthy person as sick -> after additional tests, such a case can be eliminated

Question 20	Answer - 1,3,4,6
Question 21	Answer - 2

Question #:22 What is the problem of an unbalanced dataset in binary classification?

1. This is a problem where the classes have the same frequency. For example, an image collection of 50% dogs and 50% cats.

2. This is a problem where the classes have significantly different frequencies. For example, a terrorist attack dataset where 0.01% of the examples are positive and 0.99% negative is a class imbalance problem.

Question #:23 Consider the problem of image processing. Each image is identified with a 256x256 matrix. What will be the dimension of the output matrix after applying the 5x5 kernel convolution operation?

1. 251x251

2. The matrix dimension cannot be determined from the information provided.

3. 256x256

4. 252x252

Question 22	Answer - 2
Question 23	Answer - 4

Question #:24 Select the metrics that can be used for multi-class classification problems.

1. accuracy
2. confusion matrix
3. F1 Score
4. precision
5. ROC curve

Question #:25 During binary classification problems, a confusion matrix is often built, which can be the basis for the calculation of many metrics. What is meant by True Negative (TN)?

1. True Negative is the result where the model correctly predicts the positive class.

2. True Negative is a result where the model incorrectly predicts a positive class.

3. True Negative is the result where the model correctly predicts the negative class.

4. True Negative is the result where the model incorrectly predicts a negative class.

Question 24	Answer - 1,2,3,4
Question 25	Answer - 3

Question #:26 What is the Mean Absolute Error (MAE) metric in regression problems?

1. MAE is a measure of error calculated from the mean of the absolute errors.

2. MAE is a metric that calculates the mean squared error.

3. MAE is the metric that returns the maximum residual error. This is the worst case of error between the predicted value and the true value.

Question #:27 During binary classification problems, a confusion matrix is often built, which can be the basis for the calculation of many metrics. What is meant by False Positive (FP) or Type I error?

1. False Positive is the result where the model correctly predicts the positive class.

2. False Positive is the result where the model correctly predicts the negative class.

3. False Positive is the result where the model incorrectly predicts a negative class.

4. False Positive is a result where the model incorrectly predicts a positive class.

Question 26	Answer - 1
Question 27	Answer - 4

Question #:28 What is anomaly detection?

1. The process of identifying outliers. For example, if the mean of a certain feature is 80 with a standard deviation of 10, then anomaly detection should flag a value of 150 as suspicious.

2. Anomaly detection allows you to locate the places where objects exist in the images. The appropriate algorithm places rectangular bounding boxes that fully contain the object.

Question 28	Answer - 1

Question #:29 What is BERT in machine learning terminology?

1. BERT is a metric that allows us to determine the accuracy of regression models.

2. Bit Error Rate Test

3. BERT is one of the variants of the SGD algorithm used in neural networks.

4. Model architecture for text representation. The trained BERT model can function as part of a larger text classification model or other ML tasks.

Question #:30 What is centroid (K-Means algorithm)?

1. Centroid is the manually determined primary starting point for the K-Means algorithm.

2. Centroid is the center of the cluster defined by the K-Means algorithm. For example, if k is 5, the k-means algorithm finds 5 centroids.

Question #:31 What is the AUC (Area under the ROC Curve) metric?

1. The AUC is the probability that the classifier will be more confident that a randomly selected positive example is actually positive than a randomly selected negative example is positive.

2. The AUC is the probability that the classifier will be more confident that the randomly selected negative example is actually negative than the randomly selected positive example is negative.

Question 29	Answer - 4
Question 30	Answer - 2
Question 31	Answer - 1

Question #:32 What is a convolutional neural network - CNN?

1. It is a neural network in which at least one layer is an activation layer.

2. It is a neural network in which at least one layer is a recursive layer.

3. It is a neural network in which at least one layer is a convolution layer.

Question #:33 Select metrics that can be used for binary classification problems.

1. MSE
2. precision
3. Log loss
4. F1 Score
5. confusion matrix
6. accuracy
7. ROC curve
8. MAE

Question 32	Answer - 3
Question 33	Answer - 2,3,4,5,6,7

Question #:34 What is the convolution layer in a neural network?

1. A layer of the deep neural network where the convolutional filter passes along an input matrix.

2. A layer where the activation function is applied.

3. A layer of the deep neural network in which the pooling operation is applied.

Question #:35 Select true statements about the convolution operation.

1. The term convolution in machine learning often refers to the convolutional operation or the convolutional layer in a neural network.

2. It is a combination of two functions. In machine learning, a convolution combines a convolutional filter and an input matrix to train weights.

3. Thanks to convolutions, the machine learning algorithm dramatically reduces the amount of memory needed to train the model (by reducing dimensions).

4. The use of the convolution layer of the neural network does not reduce the amount of memory needed to train the network.

Question 34	Answer - 1
Question 35	Answer - 1,2,3

Question #:36 During binary classification problems, a confusion matrix is often built, which can be the basis for the calculation of many metrics. What is meant by True Positive (TP)?

1. True Positive is the result where the model correctly predicts the negative class.

2. True Positive is the result where the model incorrectly predicts a negative class.

3. True Positive is the result where the model correctly predicts the positive class.

4. True Positive is a result where the model incorrectly predicts a positive class.

Question #:37 During binary classification problems, a confusion matrix is often built, which can be the basis for the calculation of many metrics. What is meant by False Negative (FN) or a type II error?

1. False Negative is the result where the model incorrectly predicts a negative class.

2. False Negative is the result where the model correctly predicts the positive class.

3. False Negative is a result where the model incorrectly predicts a positive class.

4. False Negative is the result where the model correctly predicts the negative class.

Question 36	Answer - 3
Question 37	Answer - 1

Question #:38 What do we mean by an agent in reinforcement learning?

1. An agent is the reward of taking action in a state determined by the environment.

2. In reinforcement learning, an agent is a world that contains the user and allows him to observe the state of this world.

3. An agent is an entity that has a policy to maximize the expected return gained from transitioning between states of the environment.

Question #:39 Consider the classic MNIST image data set (a collection of handwritten digits). Each image is a 28x28 matrix. What will be the dimension of the output matrix after applying the 3x3 kernel convolution operation?

1. 26x26

2. 31x31

3. 25x25

4. 28x28

Question 38	Answer - 3
Question 39	Answer - 1

Question #:40 What is backpropagation?

1. It is an algorithm for selecting the appropriate variables for the machine learning model.

2. The basic algorithm in cluster analysis that aims to build a hierarchy of clusters.

3. In machine learning, backpropagation is a widely used algorithm for training feedforward neural networks.

Question #:41 What is a classification model?

1. A type of machine learning model to distinguish between two or more discrete classes. For example, a natural language processing classification model can determine whether the input sentence was in English, Spanish, or Italian.

2. A type of machine learning task that predicts the continuous value of the target variable.

3. A type of machine learning task that divides data into multiple groups so that points in the same groups look more like other points in the same group than in other groups.

Question 40	Answer - 3
Question 41	Answer - 1

Module 4: Interview Preparation Part 4

Question #:1 What is the ReLU function?

1. The ReLU function is the activation function defined as follows: If the input is positive, the output is 0. If the input is non-positive, the output is equal to the input.

2. The ReLU function is the activation function defined as follows: If the input is negative, the output is 0. If the input is non-negative, the output is equal to the input.

Question #:2 Select true statements about Tensorflow.

1. Tensorflow is powered by Google.

2. Tensorflow is paid.

3. Tensorflow is an open source machine learning platform. It has a comprehensive, flexible ecosystem of tools, libraries and resources that allows you to implement the latest ML solutions.

4. Tensorflow is a distributed machine learning platform.

Question 1	Answer - 2
Question 2	Answer - 1,3,4

Question #:3 What is the interpretability of the machine learning model?

1. It is the ability to explain the ML model in an understandable way. Sometimes the models have to be interpretable (for example the right to explain - why the client did not receive the requested loan).

2. It is an inability to explain how a machine learning model works in an understandable way (black-box models).

Question #:4 Select true statements about the GPU - Graphical Processing Unit.

1. Training a neural network requires a lot of computing power, and a decent GPU ensures smooth computation of the neural networks.

2. GPU - The graphics processing unit is a single chip processor designed to efficiently manipulate computer graphics and image processing.

3. The GPU cannot be used to train neural networks.

Question 3	Answer - 1
Question 4	Answer - 1,2

Question #:5 Select true statements about transfer learning.

1. Transfer learning is the transfer of learned information from one machine learning task to another.

2. Transfer learning may involve the transfer of knowledge from the solution of a simpler task to a more complex one, or the transfer of knowledge from a task with more data to a task with much less data to learn.

3. Transfer learning is the exchange of team members between data science teams of different companies.

Question 5	Answer - 1,2

Question #:6 What is word embedding in natural language processing?

1. In natural language processing, it is a continuous sequence of n elements from a given text or speech sample. Elements can be syllables, letters, or words, for example.

2. In natural language processing, word embedding is a method of representing words for analyzing text as a real-valued vector.

3. In natural language processing, it is the process of breaking up a text string into units called tokens. The tokens can be words or a group of words. This is a key step in natural language processing.

Question #:7 What is object detection (image processing)?

1. Object detection is the segmentation of pixels into a specific category, such as a car, road, sign, or pedestrian. Image segmentation is widely used in autonomous vehicle applications to show roads, cars and people.

2. Object detection is used to classify images. Neural networks (mainly CNN) can be trained to identify different objects (classic dog vs. cat).

3. Object detection allows you to locate where objects exist. The algorithm places rectangular bounding boxes that fully contain the object. The detector can be trained to see where there are cars or people in the image.

Question 6	Answer - 2
Question 7	Answer - 3

Question #:8 Select true statements about hyperparameters.

1. Hyperparameters are set before training starts and do not change when you train the model.

2. Hyperparameters are initialized before training begins and change as you train the model.

3. Hyperparameters are model properties that allow you to control the learning process. For example, the speed at which the model can learn (learning rate) or the complexity of the model (decision tree depth, number of hidden layers in the neural network).

Question #:9 What is the loss function in neural networks (supervised ML)?

1. This is a function that is used to introduce nonlinearities into the neural network, helping it learn a more complex function.

2. It is a measure of how far a model's predictions are from its labels. The loss function measures how bad the model is.

Question 8	Answer - 1,3
Question 9	Answer - 2

Question #:10 What is JSON?

1. JavaScript Object Notation. A format for exchanging data that is structured as collections of name-value pairs or an ordered list of values.

2. JSON is a comma-separated text file format. Each line of the file is a data record. Each record consists of one or more fields separated by commas.

3. JSON, often abbreviated as JS, is a programming language that follows the ECMAScript specification. JSON is high-level, often compiled just-in-time, and multi-paradigm.

4. JSON is a semicolon delimited text file format. Each line of the file is a data record. Each record consists of one or more fields separated by commas.

Question #:11 What is image classification?

1. Image classification is used to classify images. Neural networks (mainly CNN) can be trained to identify different objects (dog vs. cat).

2. Image classification is the segmentation of pixels into a specific category, such as a car, road, sign, or pedestrian. It is widely used in autonomous vehicle applications to show roads, cars and people.

3. Image classification allows you to locate the places where objects exist. The algorithm places rectangular bounding boxes that fully contain the object. The detector can be trained to see where there are cars or people in the image.

Question 10	Answer - 1
Question 11	Answer - 1

Question #:12 What is the softmax function for?

1. The softmax function is a function that has a characteristic S-shaped curve that converts values to the range 0 to 1. It is one of the most commonly used nonlinear activation functions.

2. The softmax function provides probabilities for each possible class in a multi-class classification model. The probability adds up to exactly 1. For example, the softmax function can determine that the probability that a given image is a plane is 0.9, an auto is 0.07, and a bicycle is 0.03.

3. The softmax function is the activation function defined as follows: If the input is negative, the output is 0. If the input is non-negative, the output is equal to the input.

4. The softmax function is the activation function defined as follows: If the input is positive, the output is 0. If the input is non-positive, the output is equal to the input.

Question #:13 What is the training set for?

1. The training set is used to evaluate the predictive power of the model.

2. The training set is used to train the machine learning model.

Question 12	Answer - 2
Question 13	Answer - 2

Question #:14 What do you mean by learning rate in machine learning?

1. Learning rate is used to train the model through the gradient descent algorithm. During each iteration, the algorithm multiplies the gradient by the learning rate.

2. This is a metric showing how good the model is.

Question 14	Answer - 1

Question #:15 What is tokenization (NLP)?

1. Tokenization is the process of breaking up a text into units called tokens. The tokens can be words or a group of words. This is a key step in natural language processing.

2. In natural language processing, tokenization is a method of representing words for the analysis of text as a real-valued vector.

Question #:16 What is a recurrent neural network?

1. It is a neural network in which at least one layer is a pooling layer.

2. It is a neural network that is intentionally run multiple times, where parts of each run feed into the next run. Specifically, hidden layers from the previous run provide part of the input to the same hidden layer in the next run.

3. It is a neural network in which at least one layer is a convolution layer.

Question #:17 What is image segmentation?

1. Image segmentation is the segmentation of pixels into a specific category, such as a car, road, sign, or pedestrian. It is widely used in autonomous vehicle applications to show roads, cars and people.

2. Image segmentation enables you to locate the places where objects exist. The algorithm places rectangular bounding boxes that fully contain the object. The detector can be trained to see where there are cars or people in the image.

Question 15	Answer - 1
Question 16	Answer - 2
Question 17	Answer - 1

Question #:18 What is Keras in deep learning?

1. Python data analysis library.

2. Keras is a popular Python machine learning API. Keras runs on several deep learning frameworks, including Tensorflow, where it is made available as tf.keras.

3. A dashboard for Tensorflow that displays the summaries saved during the execution of one or more TensorFlow programs.

Question #:19 What is pooling in neural networks?

1. This is the reduction of the matrix formed by the earlier convolution layer to a matrix of a smaller size. Pooling typically involves taking the maximum or average value from the pooled area.

2. It is a combination of two functions. In machine learning, a pooling operation combines a convolution filter and an input matrix to train weights.

Question 18	Answer - 2
Question 19	Answer - 1

Question #:20 What is a tensor?

1. A dashboard that displays the summaries saved during the execution of one or more Tensorflow programs.

2. Popular Python Machine Learning API. It runs on several deep learning platforms, including Tensorflow.

3. Tensor is the basic data structure in Tensorflow. Tensors are N-dimensional data structures, most often scalars, vectors, or matrices. Tensor elements can contain integer, floating point, or string values.

Question #:21 What is a tensor?

1. Popular Python Machine Learning API. It runs on several deep learning platforms, including TensorFlow.

2. Tensor is a multidimensional array of data values, which is the basic data structure used in TensorFlow. A tensor consists of a set of values shaped into an array with any number of dimensions.

Question 20	Answer - 3
Question 21	Answer - 2

Question #:22 What does DBMS stand for in database terminology?

1. Database Management System
2. Database Main System
3. Defense-Business Management System
4. Domain Management System

Question #:23 What is automatic speech recognition (ASR)?

1. Automatic Speech Recognition takes human voice as input and converts it into human readable text. ASR allows us to create hands-free text messages and is often referred to as a speech-to-text service.

2. Automatic Speech Recognition takes text as input and converts it into synthetic speech. It is often referred to as a text-to-speech service.

Question #:24 What is optical character recognition (OCR)?

1. OCR is the conversion of images of printed, handwritten or typed text into a machine-friendly text format. OCR allows you to extract text from a scanned document.

2. OCR is the conversion of text as input into an image. This is the so-called text-to-image service.

Question 22	Answer - 1
Question 23	Answer - 1
Question 24	Answer - 1

Question #:25 What is L2 regularization?

1. A type of regularization that penalizes weights in proportion to the sum of the absolute values of the weights.

2. A type of regularization that penalizes weights in proportion to the sum of the squares of the weights. L2 regularization helps drive outlier weights closer to 0 but not quite to 0.

Question #:26 Select true statements about computer vision.

1. Computer vision can include problems such as segmentation, classification, and detection using images and movies.

2. Computer vision is primarily aimed at understanding the content of videos and photos, and then formulating useful information from them.

3. Computer vision is a branch of artificial intelligence and deep learning dedicated to developing human-like vision.

Question 25	Answer - 2
Question 26	Answer - 1,2,3

Question #:27 What does SQL stand for in database terminology?

1. Simulated Quest For Learning
2. Structured Question Line
3. Standard Question List
4. Standard Query List
5. SQL is an acronym for Structured Query Language. The SQL language is used to communicate with the database.

Question #:28 What is the input layer of the neural network?

1. This is the last layer of the neural network.
2. The input layer is also known as the hidden layer of the neural network.
3. This is the first layer in the neural network that receives input data.

Question 27	Answer - 5
Question 28	Answer – 3

Question #:29 What statement is used to commit changes made to a transaction?

1. MERGE
2. ROLLBACK
3. SAVE
4. SAVEPOINT
5. COMMIT

Question #:30 Can a table contain more than one PRIMARY KEY constraint?

1. No
2. Yes

Question 29	Answer - 5
Question 30	Answer - 1

Question #:31 What does RDBMS stand for in database terminology?

1. Relational Database Main System
2. Relational Database Management System
3. Raw Database Management System
4. Non-Relational Database Management System
5. Remote Database Management System

Question #:32 What is NER - Named Entity Recognition about in natural language processing?

1. NER is the process of breaking up a text into units called tokens. The tokens can be words or a group of words. This is a key step in natural language processing.

2. NER is an information extraction task to identify and classify named entities in the text into predetermined categories such as famous people, locations, organizations, events, etc.

3. In natural language processing, NER is a method of representing words for text analysis as a real-valued vector.

Question 31	Answer - 2
Question 32	Answer - 2

Question #:33 What is the principle of 'garbage in, garbage out' in data science?

1. The rule is that when we pass wrong input data to the model it can lead to misleading results and generate nonsensical output.

2. The rule that says that we can throw any prepared input data into the neural network, because the neural network is intelligent enough to deal with such a problem as well.

Question #:34 What is the test set for?

1. The test set is used to train the machine learning model.

2. The test set is used to evaluate the predictive power of the model.

Question #:35 Select popular optimizers used when training neural networks.

1. Adam

2. Keras

3. Adagrad

4. RMSProp

5. SGD

Question 33	Answer - 1
Question 34	Answer - 2
Question 35	Answer - 1,3,4,5

Question #:36 What is Tensorboard?

1. Popular Python Machine Learning API. It runs on several deep learning platforms, including Tensorflow.

2. Basic data structure in Tensorflow. Tensors are n-dimensional data structures, most often scalars, vectors, or matrices. Tensor elements can contain integer, floating point, or text values.

3. A dashboard that displays the summaries saved during the execution of one or more Tensorflow programs.

Question #:37 The relational database is:
a database that stores objects as opposed to data organized as tables.

1. a database in which data is divided into smaller, logically separated parts (tables) connected with each other by relations.

2. a database that stores data in the form of key-value pairs.

3. a database that can handle huge amounts of rapidly changing, unstructured data.

Question 36	Answer - 3
Question 37	Answer - 1

Question #:38 What is L1 regularization?

1. A type of regularization that penalizes weights in proportion to the sum of the absolute values of the weights.

2. A type of regularization that penalizes weights in proportion to the sum of the squares of the weights. L1 regularization helps drive outlier weights closer to 0 but not quite to 0.

Question #:39 What is epoch in neural network terminology?

1. The epoch describes how many times the learning algorithm sees the entire data set. It is one pass of all training data through the network. For example, an epoch set to 20 means the training data runs 20 times through the neural network during training.

2. It is a dataset used in one iteration (i.e., one gradient update) of training the neural network.

Question #:40 What is Markov Chain?

1. The Markov chain is a stochastic model that describes a sequence of possible events where the probability of each event depends only on the state reached in the previous event.

2. The Markov chain is a combination of the predictions of several underlying estimators to improve the generalization of the machine learning model.

Question 38	Answer - 1
Question 39	Answer - 1
Question 40	Answer - 1

Question #:41 What is the dropout layer for in neural networks?

1. The dropout layer takes the output of the previous layer's activation and randomly sets a fraction (dropout rate) of activation to 0. This is a common regularization technique used to prevent overfitting in neural networks.

2. This is the layer of the neural network that applies the activation function to the output.

3. The layer of the neural network where the convolutional filter is applied.

Question #:42 What is Anaconda in data science?

1. Anaconda is an open-source platform for the Python and R programming languages that aims to simplify the management of environments, packages, and deployment. The package versions in Anaconda are managed by the conda package management system.

2. Anaconda is the core Python package for data analysis.

3. It's a popular Python Machine Learning API. Anaconda runs on several deep learning platforms, including TensorFlow.

Question 41	Answer - 1
Question 42	Answer - 1

Question #:43 Select true statements about sentiment analysis.

1. Sentiment analysis uses machine learning algorithms to determine a group's overall attitude - positive, neutral or negative - to a service, product, organization, or topic.

2. An example of sentiment analysis is the use of natural language understanding to analyze sentiment towards a company based on stock market information related to that company.

Question 43	Answer - 1,2

Module 6: Interview Preparation Part 5

Question #:1 What is normalization in database terminology?

1. Normalization is a data storage strategy that duplicates data in different tables, rather than joining tables with foreign keys and join queries. Database normalization avoids costly joining operations.

2. Database normalization is the process of structuring a database, in accordance with a series of so-called normal forms in order to reduce data redundancy and improve data integrity.

Question #:2 Select the types of relationships in relational databases.

1. One-to-Many
2. One-to-Zero
3. Many-to-Many
4. One-to-One
5. One-to-Three
6. One-to-Two

Question 1	Answer - 2
Question 2	Answer - 1,3,4

Question #:3 What is the difference between a view and a table?

1. The view is where the database actually stores its data. The view takes up disk space and is created by defining columns, data types, and constraints.

2. A tabel is a virtual entity whose content is defined by the query.

3. The table is where the database actually stores its data. The table takes up disk space and is created by defining columns, data types, and constraints.

4. A view is a virtual table whose content is defined by the query.

Question #:4 What is a database?

1. A database is an organized collection of data. The data is stored according to certain rules, usually stored electronically. A database management system (DBMS) is used to work with the database. Data, a database management system, and related applications together form a database system, which we often refer to as a database for short. Most often, the data in the database is stored in the form of tables.

2. A database is a collection of data stored electronically in an unorganized manner.

Question 3	Answer - 4
Question 4	Answer - 1

Question #:5 Select true statements.

1. An index is a data structure that improves the speed of data retrieval operations on a database table at the cost of additional writes and storage space to maintain the index data structure.

2. Indexes can be created using one or more columns of a table.

3. Indexes are used to quickly locate data without having to search every row in a database table every time a database table is accessed

Question #:6 Select true statements (DDL).

1. DDL is a syntax for creating and modifying database objects such as tables, indices, and users.

2. With the DDL statement, a user does not directly manipulate the data, but its structure.

3. Common examples of DDL statements include CREATE, ALTER, and DROP.

Question 5	Answer - 1,2,3
Question 6	Answer - 1,2,3

Question #:7 What is a CROSS JOIN?

1. A CROSS JOIN is a regular join, but the table is joined with itself. It is useful when we want to combine pairs of rows from the same table.

2. The CROSS JOIN keyword returns all records from the left table, and the matching records from the right table.

3. The CROSS JOIN selects records that have matching values in one or more tables.

4. The CROSS JOIN produces a result set which is the number of rows in the first table multiplied by the number of rows in the second table. This kind of result is called as Cartesian Product. We need to be careful with CROSS JOINs because the query result can be quite large.

Question #:8 What is SQL INJECTION?

1. SQL injection is a code injection technique used to attack data-driven applications, in which malicious SQL statements are inserted into an entry field for execution. It might destroy your database.

2. It is an instruction that allows you to insert data into several tables at the same time.

Question 7	Answer - 4
Question 8	Answer - 1

Question #:9 Select true statements (primary key).

1. The primary key can consist of one or more columns.

2. A primary key cannot consist of multiple columns.

3. The PRIMARY KEY constraint uniquely identifies each record in the table. Primary keys must contain UNIQUE values and must not contain NULL values.

4. A table can only have one primary key.

Question #:10 Select correct statements (data warehouse).

1. Data warehouses enable enterprises to extract valuable business information that simplifies the decision-making process.

2. A data warehouse is a tool that collects and stores large amounts of data in one place, coming from various, often dispersed sources.

3. Data warehouses are primarily intended for query (read-only) and analytics.

4. The main purpose of a data warehouse is to enable business analysis (BI) and various types of analytics.

5. Data warehouses are mainly designed to support transaction databases.

Question 9	Answer - 1,3,4
Question 10	Answer - 1,2,3,4

Question #:11 What does the Consistency property in ACID mean?

1. The property of a transaction that guarantees that the changes made by a transaction are isolated from the rest of the system until after the transaction has committed.

2. The property of a transaction that guarantees that the state of the database both before and after execution of the transaction remains consistent whether or not the transaction commits or is rolled back.

3. The property of a transaction in which the DBMS guarantees that all committed transactions will survive any kind of system failure.

4. The property of a transaction that guarantees that either all or none of the changes made by the transaction are written to the database.

Question #:12 What is DML - Data Manipulation Language?

1. DML - Data Manipulation Language is a group of SQL statements used for inserting, deleting, and modifying data in a database.

2. DML is a group of SQL statements for creating and modifying database objects such as tables, indices, and users.

3. DML statements are used for performing queries on the data.

Question 11	Answer - 2
Question 12	Answer - 1

Question #:13 What is DCL - Data Control Language?

1. DCL is a group of SQL statements that define data structures.

2. DCL is a group of SQL statements that are used to modify, insert, and delete data.

3. DCL is used to control access to data stored in a database.

Question #:14 What is the DEFAULT constraint for?

1. The DEFAULT constraint is used to set the default table.

2. The DEFAULT constraint is used to populate a column with a default value.

3. The DEFAULT constraint is used to set the default database.

4. The DEFAULT constraint is used to set the default user.

Question #:15 What is ETL?

1. ETL is a type of data integration that refers to the three steps (extract, transform, load) used to blend data from multiple sources. During this process, data is(extracted from a source system, transformed into a format that can be analyzed, and loaded into a data warehouse or other system.

2. ETL can be described as a database backup process.

Question 13	Answer - 3
Question 14	Answer - 2
Question 15	Answer - 1

Question #:16 Select true statements (triggers).

1. A trigger is procedural code that is automatically executed in response to certain events on a particular table or view in a database.

2. There are several types of triggers. BEFORE triggers - executed before the statement generating the event. AFTER triggers are executed after the instruction that generates the event. Some databases also have INSTEAD OF triggers - these are executed instead of the instruction that generates the event.

3. The trigger is mostly used for maintaining the integrity of the information on the database.

4. A trigger is a data structure that improves the speed of data retrieval operations on a database table at the cost of additional writes and storage space to maintain the index data structure.

Question 16	Answer - 1,2,3

Question #:17 Select true statements (foreign key).

1. Foreign key is a type of constraint that maintains the consistency of the database. This constraint can prevent data from being inserted or updated if the data becomes inconsistent.

2. A table can contain multiple columns with a foreign key constraint.

3. A table can only have one column with a foreign key constraint.

4. When a DML operation is performed, foreign key constraints can delete the data in the child tables, change it to something else, or set it to NULL based on the ON CASCADE option specified when creating the foreign key.

5. A column is a foreign key for a given table if it is not its primary key but its values are the primary key values of another table.

Question 17	Answer - 1,2,4,5

Question #:18 What does ACID stand for?

1. Accelerated Change In Design
2. Application Context Identifier
3. Atomicity, Consistency, Isolation, Durability
4. Atomicity, Consistency, Identity

Question #:19 What is OLTP?

1. Database system for performing multidimensional analyzes at high speed on large amounts of data.

2. A database system or database application that runs a workload with multiple transactions, with frequent writes and reads, typically affecting small amounts of data at once.

Question #:20 Select DCL statements.

1. GRANT
2. REVOKE
3. CONNECT
4. ACCESS

Question 18	Answer - 3
Question 19	Answer - 2
Question 20	Answer - 1,2

Question #:21 What is the Turing Test?

1. It is a test developed by Alan Turing in the 1950s that tests the ability of a machine to imitate human behavior. It was designed to determine whether a computer can be classified as intelligent.

2. The Turing test is a methodology for testing user experience. Such tests consist of a randomized experiment with two variants, A and B. Tests are a way to compare two versions of one variable, usually by testing the user's response to variant A with variant B and determining which of the two variants is more effective.

Question #:22 What is the difference between the WHERE and HAVING clauses?

1. A HAVING clause is like a WHERE clause, but applies only to groups as a whole, whereas the WHERE clause applies to individual rows. A query can contain both a WHERE clause and a HAVING clause.

2. A WHERE clause is like a HAVING clause, but applies only to groups as a whole, whereas the HAVING clause applies to individual rows. A query can contain both a WHERE clause and a HAVING clause.

Question 21	Answer - 1
Question 22	Answer - 1

Question #:23 Select true statements (constraints).

1. Constraints allow you to specify the type of data that can be inserted into a given column.

2. Constraints could be either on a column level or a table level.

3. Constraints are a key component of the ACID philosophy.

4. Constraints are the rules enforced on the data columns of a table. These are used to limit the type of data that can go into a table.

Question 23	Answer - 1,2,3,4

Question #:24 What does the Atomic property in ACID mean?

1. The property of a transaction that guarantees that the changes made by a transaction are isolated from the rest of the system until after the transaction has committed.

2. The property of a transaction in which the DBMS guarantees that all committed transactions will survive any kind of system failure.

3. The property of a transaction that guarantees that the state of the database both before and after execution of the transaction remains consistent whether or not the transaction commits or is rolled back.

4. The property of a transaction that guarantees that either all or none of the changes made by the transaction are written to the database.

Question #:25 What does the Durability property in ACID mean?

1. The property of a transaction in which the DBMS guarantees that all committed transactions will survive any kind of system failure.

2. The property of a transaction in which the DBMS guarantees that all committed transactions will survive any kind of system failure.

3. The property of a transaction that guarantees that the state of the database both before and after execution of the transaction remains consistent whether or not the transaction commits or is rolled back.

4. The property of a transaction that guarantees that either all or none of the changes made by the transaction are written to the database.

Question 24	Answer - 4
Question 25	Answer - 1

Question #:26 What does OLAP stand for?

1. Online Analytical Processing
2. Online Legal Access Project
3. Online App
4. Online Access Program

Question #:27 What is a SELF JOIN?

1. The SELF JOIN selects records that have matching values in one or more tables.

2. A self join is a regular join, but the table is joined with itself. It is useful when we want to combine pairs of rows from the same table.

3. The SELF JOIN keyword returns all records from the left table, and the matching records from the right table.

4. The SELF JOIN produces a result set which is the number of rows in the first table multiplied by the number of rows in the second table. This kind of result is called as Cartesian Product.

Question 26	Answer - 1
Question 27	Answer - 2

Question #:28 What is the WHERE clause for?

1. The WHERE clause is only used in conjunction with the GROUP BY clause to filter groups of records.

2. The WHERE clause is used to filter records. It is used to extract only those records that fulfill a specified condition.

Question #:29 What statement is used to commit changes made to a transaction?

1. MERGE

2. ROLLBACK

3. SAVE

4. SAVEPOINT

5. COMMIT

Question 28	Answer - 2
Question 29	Answer - 5

Question #:30 Can a table contain more than one PRIMARY KEY constraint?

1. No
2. Yes

Question 30	Answer - 1

Question #:31 What is the difference between OLAP and OLTP?

1. The main difference between these two systems lies in their names: analytical vs. transactional. Each system is optimized for the appropriate type of processing.

2. OLTP is optimized to perform complex data analysis to make smarter business decisions. On the other hand, OLAP is optimized to process a huge number of transactions.

3. The main difference between these two systems lies in their names: analytical vs. transactional. Each system is optimized for the appropriate type of processing.

4. OLAP is optimized to perform complex data analysis to make smarter business decisions. On the other hand, OLTP is optimized to process a huge number of transactions.

Question #:32 What is the difference between OLAP and OLTP?

1. The main difference between these two systems lies in their names: analytical vs. transactional. Each system is optimized for the appropriate type of processing.

2. OLAP is optimized to perform complex data analysis to make smarter business decisions. On the other hand, OLTP is optimized to process a huge number of transactions.

3. The main difference between these two systems lies in their names: analytical vs. transactional. Each system is optimized for the appropriate type of processing.

4. OLTP is optimized to perform complex data analysis to make smarter business decisions. On the other hand, OLAP is optimized to process a huge number of transactions.

Question 31	Answer - 4
Question 32	Answer - 2

Question #:33 What does the Isolation property in ACID mean?

1. The property of a transaction in which the DBMS guarantees that all committed transactions will survive any kind of system failure.

2. The property of a transaction that guarantees that the state of the database both before and after execution of the transaction remains consistent whether or not the transaction commits or is rolled back.

3. The property of a transaction that guarantees that the changes made by a transaction are isolated from the rest of the system until after the transaction has committed.

4. The property of a transaction that guarantees that either all or none of the changes made by the transaction are written to the database.

Question #:34 What is a transaction?

1. In a database management system, a transaction is a single unit of logic or work, sometimes made up of multiple operations.

2. A transaction is always just a single database operation.

Question #:35 Can a table have multiple foreign keys?

1. No

2. Yes

Question 33	Answer - 3
Question 34	Answer - 1
Question 35	Answer - 2

Question #:36 Select true statements (views).

1. View is another name for a table.

2. In SQL, a view is a virtual table based on the result-set of an SQL statement.

3. A view contains rows and columns, just like a real table. The fields in a view are fields from one or more real tables in the database.

4. You can add SQL statements and functions to a view and present the data as if the data were coming from one single table.

5. Views can be used to tailor the database to the needs of different groups of users. They provide a view from which a given class of users can see the database. Different groups of users may have different views of the data in the database.

Question #:37 What does ETL stand for?

1. Educational Technologies Limited

2. Endorsed Tools List

3. Extract-Transform-Load

4. Embedded Test Language

5. Extended Trading Line

Question 36	Answer - 2,3,4,5
Question 37	Answer - 3

Question #:38 What are indexes in a database for?

1. Indexes allow us to quickly retrieve records from a database.

2. Indexes are necessary when we want to join tables.

Question #:39 Select true statements (CHECK constraint).

1. The CHECK constraint is used to check user permissions to perform database operations.

2. The CHECK constraint is used to limit the range of values that can be placed in a column.

3. If we define a CHECK constraint at the column level, the constraint will only allow the specified values for that column.

4. If we define a CHECK constraint at the table level, we can restrict the column values based on the values in any of the table columns.

Question #:40 What is a one-to-one relationship?

1. A one-to-one relationship occurs when multiple records in a table are associated with multiple records in another table.

2. In a one-to-one relationship, one record in a table is associated with one and only one record in another table.

3. In a one-to-one relationship, one record in a table can be associated with one or more records in another table.

Question 38	Answer - 1
Question 39	Answer - 2,3,4
Question 40	Answer - 2

Question #:41 What is the UNION operator for?

1. The UNION operator is used to join columns in a table.

2. The UNION operator is used to concatenate the result set of two or more SELECT statements.

3. The UNION operator does not exist in SQL.

Question 41	Answer - 2

ABOUT THE AUTHOR

“Edcorner Learning” and have a significant number of students on **Udemy** with more than **90000+ Student and Rating of 4.1 or above.**

Edcorner Learning is Part of Edcredibly.

Edcredibly is an online eLearning platform provides Courses on all trending technologies that maximizes learning outcomes and career opportunity for professionals and as well as students. Edcredibly have a significant number of 100000+ students on their own platform and have a **Rating of 4.9 on Google Play Store – Edcredibly App**.

Feel Free to check or join our courses on:

Edcredibly Website - https://www.edcredibly.com/

Edcredibly App –
https://play.google.com/store/apps/details?id=com.edcredibly.courses

Edcorner Learning Udemy -
https://www.udemy.com/user/edcorner/

Do check our other eBooks available on Kindle Store.