Ques 1 = Glm is used to describe the relation between two variables and thus deciding whether that relation is statistically significant or not. The specific version of the glm we use is referred to as linear regression.

Ques2 = it has four key assumptions i.e linearity, constant variance, normality,independence.

Ques 3 = interpreting the coefficients in a generalized linear model depends on the specific type of glm. In glm coefficients represent the changes involved in the variable which is being associated with a one unit change in a predicted value, while keeping the other value constant.

Ques 4 = it all depends on the variable number that is being considered in the model.

1. Univariate GLM = univariate glm focuses on analyzing and modeling only one response at a time.it accesses the relationship b/w one response variable and one predicted value.

2. Multivariate GLM =multivariate GLM focuses on analyzing and modeling multiple response variables simultaneously. On the other hand it gives us permission to examine the relationship b/w multiple response variables and predicted values.

Ques 5 = in generalized linear model interaction effects when the relationship b/w the predicted value and response variable present or the level of different predictor value. It shows that the predicted value effect on the response is not constant.

Ques 6 = we handle categorical predictors by converting them into numerical variables.

1. Dummy coding
2. Effect coding
3. Polynomial coding

It is crucial to note that the coding scheme chosen may be influenced by the study topic, the nature of the categorical variable, and the statistical programme employed. Furthermore, it is critical to accurately interpret the coefficients and hypothesis tests depending on the chosen coding scheme.

Ques 7 = The model matrix is also known as predictor matrix which is also the fundamental component of GLM. it is also used in organizing other stuff and presenting the predicted variables that are used to estimate the model parameter and thus end up in making the decisions.
It plays a vital role in making the data transformed the raw data into a format that can be analyzed and interpreted within the GLM

Ques 8 = In a Generalized Linear Model (GLM), the significance of predictors can be tested using hypothesis tests, specifically through the p values associated with the coefficients.if the p value is small it tells us that the predictor values is directly associated with the response variable.

Ques 9 = In the context of Generalized Linear Models (GLMs), Type I, Type II, and Type III sums of squares refer to different approaches for partitioning the variability in the response variable among the predictor variables. It is important to note that different software packages and statistical procedures may use different default settings for calculating sums of squares, so it's essential to understand the specific implementation used in your chosen software or statistical method.

Ques 10 = deviance in a GLM provides a measure of the lack of fit between the observed data and the fitted model. Minimizing the deviance is the primary objective in GLM analysis, and it serves as the basis for goodness-of-fit assessment, hypothesis testing, and model comparison.

Ques 11
It is a statistical method used to investigate the relationship between a dependent and independent variable. It is also used in prediction, explanation and inference. It also helps to draw meaningful insights from the data.

Ques 12
Main difference between multiple regression and linear regression is that no of variables present and used in the model the relationship with the dependent var.

1. No. of predictors
2. 2.complexity
3. Model interpretation
4. Model evaluation

Ques 13
It is also known as the coefficient of determination . it gives us the evaluation that the model is good fit. It also ranges between 0 to 1
Where 0 = The model explains none of the variability in the dependent variable.
And 1 = The model explains all of the variability in the dependent variable.

Interpreting the R-squared value involves considering its magnitude and contextualizing it within the specific analysis context

Ques 14
Correlation and regression are both statistical approaches used to analyze the connection between variables, but they differ in their aims, measurements, and kind of study.
Correlation assesses the degree and direction of a linear relationship between variables, whereas regression attempts to model and forecast the dependent variable using the independent variables.

Ques 15
The coefficients and intercept are crucial components of the regression model in regression analysis because they reflect the connection between the independent variables and the dependent variable.
In a regression model, the coefficients and the intercept work together to represent the connection between the independent variables and the dependent variable. Based on the values of the independent variables, they jointly predict the values of the dependent variable.

Ques 16
Handling outliers in regression analysis is a key step in ensuring the results' robustness and reproducibility. Outliers are data points that differ greatly from the main trend of the data and have a disproportionate effect on the regression model.

1. Identify outlier
2. Nature of outlier
3. Impact of outlier
4. Data transformation
5. Remove outlier
6. Robust regression

QUes 17
Ridge regression and ordinary least squares (OLS) regression are both regression techniques used to model the connection between dependent and independent variables. Their techniques for dealing with multicollinearity and the influence of predictor factors, however, differ.The choice between ridge regression and OLS regression depends on the presence of multicollinearity and the goals of the analysis.

Ques 18

In regression, heteroscedasticity refers to a failure to meet the assumption that the variability of the mistakes (residuals) is consistent across different levels of the independent variables. In other words, it happens when the spread or dispersion of residuals is inconsistent over the anticipated value range.

Ques 19

Multicollinearity occurs when there is a high correlation or linear relationship among the independent variables in a regression model. It can lead to issues in regression analysis, including unstable coefficient estimates, difficulties in interpreting the individual effects of predictors, and reduced predictive accuracy. It is vital to remember that no single method is generally applicable, and the methodology chosen relies on the individual context, research objective, and data type. Additionally, talking with a statistician or subject-matter expert can give useful insights for properly dealing with multicollinearity.

Ques 20

Polynomial regression is a form of regression analysis in which the relationship between the independent variable(s) and the dependent variable is modeled as an nth-degree polynomial function. Unlike linear regression, which assumes a linear relationship between the variables, polynomial regression allows for capturing nonlinear relationships between the variables.

1. Nonlinear relationship
2. Overfitting
3. Exploratory analysis

Ques 21

A loss function measures the difference between predicted and target values, drives the learning process by optimizing model parameters, aids in model performance evaluation, and allows model selection and comparison. It is critical in machine learning because it drives the optimisation and training of models to increase their capacity to generate correct predictions.

Ques 22

Non-convex loss functions can have several local minima and are difficult to optimize, whereas convex loss functions have a single global minimum and are easy to optimize. Non-convex loss functions are more prevalent in complicated models and tasks, but

they require careful initialization and optimization procedures to discover appropriate solutions.

Ques 23
In a regression issue, mean squared error measures the average squared difference between predicted and true values. It is derived by averaging the squared residuals and offers a measure of the overall size of prediction mistakes.

Ques 24
In a regression issue, mean absolute error measures the average absolute difference between predicted and true values. It is generated by averaging the absolute residuals and offers a measure of the average size of forecast mistakes. MAE is a valuable assessment metric for regression models since it is easily interpretable and resistant to outliers.

Ques 25
Log loss (cross-entropy loss) is a classification loss function that measures the dissimilarity between predicted and true class labels. It is derived by adding the logarithmic losses for all observations and average them. Because of its sensitivity to prediction confidence and usefulness in probabilistic classification tasks, log loss is commonly used.

Ques 26
The nature of the issue, the kind of job (regression or classification), the qualities of the data, and the evaluation metrics of interest all influence the choice of loss function for a specific problem.

1. Problem type
    1. Regression
    2. Classification

2. Data characteristics
    1. Imbalance
    2.noisy

3. Task specific
    1.taks requirement
    2.domain knowledge

Ques 27
Regularization is a machine learning approach used to reduce overfitting and enhance model generalization performance. It entails altering the loss function by including a regularization term that imposes a penalty on the complexity or size of the model's parameters. The goal of regularization is to strike a compromise between fitting the training data effectively and keeping things simple or avoiding overly complex parameter values.

There are two regularization

1. Ridge (L2)
2. Lasso(L1)

Ques 28
The Huber loss function combines the properties of mean squared error (MSE) and mean absolute error (MAE). It is frequently employed in regression tasks to address the impact of outliers on model performance. Huber loss strikes a compromise between outlier resistance and sensitivity to tiny mistakes.

1. Definition
2. Handling outliers

Ques 29
Quantile loss, also known as pinball loss or check loss, is a loss function used in quantile regression to model and estimate conditional quantiles of a target variable. It is particularly useful when the focus is on capturing the entire distribution of the target variable rather than just its mean.

Ques 30
The difference between squared loss and absolute loss lies in how they measure the discrepancy between predicted values and true values in a regression problem.

1. Squared loss
2. Absolute loss

Ques 31

An optimizer is an algorithm or approach in machine learning that modifies the parameters of a model to minimize the loss function and increase the model's performance. It is critical in the training phase of machine learning models since it iteratively updates the model's parameters depending on the gradient or other optimisation criteria. An optimizer's goal is to discover the best set of parameters for minimizing the gap between the model's predictions and the real values of the target variable.

Ques 32

Gradient Descent (GD) is an iterative optimization approach for minimizing a differentiable loss function. It is commonly used in machine learning to estimate parameters and train models. Gradient Descent's main notion is to iteratively update the model's parameters in the direction of the steepest descent of the loss function.

1. Objective
2. Gradient calculation
3. Parameter update
4. Iterative process
5. Optimization challenges

Ques 33

1. Batch gradient descent
2. Stochastic Gradient Descent
3. Mini batch gradient descent
4. Momentum based

Ques 34

Gradient Descent's learning rate is a hyperparameter that controls the step size or the pace at which the model's parameters are updated throughout each iteration. The magnitude of parameter updates is controlled by the gradient of the loss function. It is critical to select an adequate learning rate for successful model training and convergence.

Ques 35

Gradient Descent (GD) is susceptible to getting stuck in local optima, which are suboptimal solutions in the optimization landscape. It is crucial to highlight that GD does not always discover the global optimum, especially in non-convex situations with several local optima. The performance of GD in dealing with local optima is affected by the issue, data, initialization, learning rate, and optimisation landscape. Exploring

multiple methodologies and hyperparameter settings might increase the likelihood of discovering better solutions and reducing the influence of local optima.

Ques 36
The Stochastic Gradient Descent (SGD) optimisation method is a version of the Gradient Descent (GD) optimisation technique, which is widely employed in machine learning. SGD computes the gradient of the loss function using just one randomly picked training sample at a time, whereas GD computes the gradient utilizing the whole training dataset in each iteration.

SGD is a popular choice for working with huge datasets since it is computationally efficient and can handle flowing data. Despite introducing more noise and experiencing greater swings during training, it can nonetheless converge to a suitable solution and avoid local optima. For large-scale situations, GD delivers more exact gradient estimations but may be computationally costly.

Ques 37
The batch size of Gradient Descent (GD) optimization refers to the number of training samples utilized in each iteration to compute the gradient of the loss function. It is critical in defining the trade-off between computing efficiency and gradient estimation quality.

Ques 38
Momentum is a strategy used in optimisation algorithms such as Gradient Descent to speed the convergence process and improve optimisation efficiency. It provides a momentum term that uses past gradient knowledge to drive optimisation adjustments.

Ques 39
The decision between BGD, MBGD, and SGD is influenced by several variables, including:

Dataset Size: For big datasets, BGD may be computationally infeasible, making MBGD or SGD more appropriate.
Computational Resources: When compared to SGD, which processes one sample at a time, BGD and MBGD may demand more memory.

Smoothness vs. Noise: BGD gives the smoothest updates, but SGD introduces more noise. MBGD achieves a happy medium between the two.
SGD frequently converges quicker, although BGD and MBGD may produce more stable convergence.

Considerations for Hyperparameters: BGD requires learning rate adjustment, MBGD adds an extra mini-batch size parameter, and SGD involves learning rate tuning and maybe learning rate scheduling.

Ques 40
In Gradient Descent (GD), the learning rate is a crucial hyperparameter that governs the step size or the pace at which the model's parameters are changed throughout each iteration. It is very important in defining the convergence behavior of GD.

Ques 41
Regularization is a machine learning approach used to reduce overfitting and enhance model generalization performance. When a model grows too complicated, it begins to memorize the training data rather than understanding the underlying patterns. As a result, while the model may perform well on training data, it may fail to generalize adequately to new data.
Regularization adds limitations or penalties to the learning algorithm in order to prevent overly complicated models. It seeks to strike a compromise between fitting the training data effectively and being simple.

Regularization makes machine learning models more stable and less prone to overfitting.

Ques 42
L1 and L2 regularization are two popular regularization techniques used in machine learning. They differ in the type of penalty applied to the model's weights.
When feature selection or sparsity is critical, L1 regularization is generally used, but L2 regularization is usually used for overall weight regularization without completely deleting features. In certain circumstances, a hybrid of the two (known as elastic net regularization) can be employed to capitalize on the advantages of both procedures.

Ques 43
Ridge regression is a technique for regularizing linear regression models. It is a kind of L2 regularization that helps to prevent overfitting and enhances the model's generalization performance. Ridge regression does this by adding a penalty term to the linear regression objective function depending on the sum of the squared weights.

Ques 44
Elastic Net regularization is a regularization approach that incorporates both L1 (Lasso) and L2 (Ridge) penalties. It is intended to circumvent some of the constraints of individual regularization approaches by providing a flexible method for handling feature selection and model regularization at the same time.

Ques 45
Regularization helps prevent overfitting in machine learning models by introducing additional constraints or penalties that discourage overly complex models. Overfitting occurs when a model becomes too specialized in fitting the training data and fails to generalize well to unseen data.

Ques 46
Early stopping is a machine learning strategy that prevents overfitting by terminating the training process before the model begins to overfit the data. It is a type of regularization in which the model's performance on a validation set is monitored throughout training and training is stopped when the performance begins to deteriorate.

Ques 47
Dropout regularization is a neural network strategy for preventing overfitting and improving generalization performance. During the training phase, a proportion of neurons are randomly dropped out (deactivated).Each neuron in a neural network accepts inputs, applies an activation function, and outputs an output. Dropout regularization randomly sets a percentage of the neurons' outputs to zero during training. That is, those neurons are essentially "dropped out" or deactivated for that specific training example. The procedure is performed for each training example and iteration (epoch).

Ques 48
Choosing the regularization parameter in a model is an important task that requires careful consideration. The regularization parameter controls the amount of

regularization applied to the model, striking a balance between fitting the training data well and preventing overfitting. The specific method for choosing the regularization parameter depends on the algorithm and the problem at hand.

1. Grid search cv
2. Cross validation
3. Optimization
4. Information criteria
5. Domain knowledge

Ques 49

Before training the model, feature selection focuses on finding and choosing a subset of important characteristics with the goal of reducing dimensionality and improving interpretability. Regularisation, on the other hand, is a training approach used to limit model complexity, reduce overfitting, and increase generalization. While both deal with the issue of model complexity, feature selection works on the feature level, whereas regularization works on the model's parameters or weights.

Ques 50

There is a trade-off between bias and variance in regularized models. Bias refers to the mistake generated in the learning method by simplifying assumptions, whereas variance relates to the model's sensitivity to variations in the training data. It is critical to balance bias and variance in order to get excellent generalization performance.

Ques 51

SVM is a supervised machine learning technique that is used for classification and regression applications. It is especially useful for tackling binary classification issues, which require categorizing data points into two groups.

Support Vector Machines have shown to be extremely effective classifiers, especially when working with structured and high-dimensional data. Because of their capacity to establish optimal decision limits and deal with nonlinear interactions, they are frequently employed in a variety of applications such as picture classification, text categorization, and bioinformatics.

Ques 52

SVM's kernel technique enables nonlinear classification by implicitly translating data points into a higher-dimensional feature space via a kernel function. Even when the original data points are not linearly separable, SVM can discover optimum linear decision boundaries in the converted space.

Ques 53
Support vectors in Support Vector Machines (SVM) are the data points from the training set that are closest to the decision border (hyperplane). These support vectors are crucial in defining the decision boundary and finding the SVM model's best solution.

Ques 54
In SVM, the margin reflects the distance between the decision border and the support vectors. In general, a higher margin leads to greater generalization, enhanced resilience to outliers, and improved model performance. It enables SVM to strike a balance between fitting the training data and avoiding overfitting, resulting in more trustworthy predictions on previously unknown data.

Ques 55
Handling unbalanced datasets in SVM can be important when the number of data points in one class is significantly higher than the other class. SVM's performance can be affected in such cases, as it tends to prioritize the majority class and may struggle to classify the minority class accurately.

1. Adjust class weights
2. Undersampling
3. Oversampling
4. Hybrid approaches

Ques 56
The capacity to manage the complexity of decision boundaries distinguishes linear SVM from non-linear SVM. Linear SVM finds a straight-line or hyperplane decision boundary for linearly separable data, but non-linear SVM employs the kernel method to convert the data into a higher-dimensional space to handle nonlinear interactions and discover more complicated decision boundaries. Non-linear SVM is more adaptable and can

handle a wider range of classification tasks, although it may be more computationally demanding than linear SVM. The decision between linear SVM and non-linear SVM is dictated by the nature of the data and the difficulty of the classification problem at hand.

Ques 57

The C-parameter, also known as the regularization parameter, is a crucial hyperparameter in Support Vector Machines (SVM). It influences the SVM model's handling of misclassified data points and the trade-off between maximizing the margin and minimizing the classification error. The C-parameter controls the balance between achieving a wider margin and allowing margin violations.

1. Margin violations and misclassification
2. Model complexity and overfitting
3. Sensitive to outliers

Ques 58

In Support Vector Machines (SVM), slack variables are introduced to handle cases where the data points are not perfectly separable by a hyperplane. These variables allow SVM to handle margin violations and misclassified instances while still aiming to maximize the margin and find an optimal decision boundary.

Ques 59

When the data is fully separable and there are no misclassifications or margin violations, hard margin SVM is utilized. When there is noise or misclassification, soft margin SVM is used to create a compromise between maximizing the margin and dealing with the defects in the data.

Ques 60

Understanding the underlying patterns and factors impacting the model's predictions is aided by comprehending the coefficients in an SVM model.

Ques 61

A decision tree is a type of supervised machine learning algorithm used for classification and regression applications. Each internal node represents a feature or property, each branch represents a decision or rule based on that characteristic, and each leaf node represents the outcome or prediction. Decision trees are named "trees"

because their structure is hierarchical, with the root node at the top and the leaf nodes at the bottom.

1. Data preparation
2. Feature selection
3. Splitting
4. Stopping criteria
5. Leaf node assignments
6. predictions

Ques 62

Splits are used in a decision tree to partition data depending on the values of a given feature or attribute. The objective is to identify the feature and associated splitting criterion that optimally divides the data into various groups or lowers impurity within each partition.

Ques 63

In decision trees, impurity measurements such as the Gini index and entropy are used to evaluate the quality of a split and select which feature to utilize for dividing the data. These measurements quantify impurity or disorder within a set of target values and are used to analyze the homogeneity or purity of subsets formed by a split.

The Gini index is a measure of impurity that is often employed in classification difficulties. It computes the likelihood of misclassifying a randomly selected member from the set. The Gini index has a value between 0 and 1, with 0 indicating pure or homogeneous subsets and 1 indicating greatest impurity or equal distribution across classes.

Another impurity metric that is widely used in decision trees and classification issues is entropy. It computes the average amount of information or uncertainty required to categorize a set element. Entropy can range from 0 and log(K), where K is the number of classes.

Ques 64

In a decision tree, information gain is used to assess the performance of several features in lowering entropy and selecting the most informative feature for dividing the data. It aids in the construction of a tree that maximizes class separation and increases the predictive capability of the model.

Ques 65

1. Ignoring the missing values
2. Missing value as a separate category
3. Imputation
4. Algo specific handling

Ques 66

Pruning is a technique used to minimize the complexity of decision trees by removing unneeded branches or nodes. It aids in the prevention of overfitting, which occurs when the decision tree becomes overly specialized to the training data and performs badly on fresh, previously unknown data. Pruning is necessary for a number of reasons:

1. overfitting prevention
2. Complexity reduction
3. Computational efficiency

Ques 67

Classification trees and regression trees have comparable construction and splitting criteria, but the evaluation of impurity or split quality differs. Impurity metrics such as the Gini index or entropy are used in classification trees to assess the purity of class assignments, whereas mean squared error (MSE) or mean absolute error (MAE) are commonly employed in regression trees to assess the accuracy of projected values.

Ques 68

Interpreting decision boundaries in a decision tree involves understanding how the tree partitions the feature space to make predictions or assign classes. Decision boundaries are the boundaries or regions where the decision tree assigns different outcomes or classes to instances based on their feature values.

Ques 69

1. Feature selection
2. Feature ranking
3. Interpretability
4. Feature engineering

Ques 70

Ensemble approaches are machine learning methods that integrate numerous separate models, known as base learners, to increase predictability and robustness. These models are trained individually, and their predictions are then integrated to provide final predictions or classifications. Because of their versatility and capacity to capture complicated interactions, decision trees are frequently employed as base learners in ensemble approaches.

Ques 71

In machine learning, ensemble approaches combine numerous models to increase predictive performance and create more accurate predictions than any one model could achieve alone. The theory behind ensemble approaches is that by combining numerous models' predictions, the strengths of each individual model may compensate for its deficiencies, resulting in a more robust and accurate forecast.

Ques 72

Bagging, short for Bootstrap Aggregating, is a machine learning ensemble approach intended to increase model accuracy and stability. It entails randomly selecting with replacement to create numerous subsets of the original training dataset and training a different model on each subset. These models' forecasts are then pooled to get the final prediction.

1. Bootstrap sampling
2. Model training
3. Prediction aggregation

Ques 73

Bootstrapping is a sampling technique used in bagging to construct numerous subsets of the original training dataset. It entails randomly choosing instances from the dataset

and replacing them with new examples to produce new subsets of the same size as the original dataset.
How it works

1. Original training dataset
2. Random sampling with replacement
3. Model training on subsets

Ques 74
Boosting is a machine learning ensemble strategy that combines numerous weak models to build a strong model. Unlike bagging, which trains models separately, boosting trains models consecutively, with each model seeking to fix the faults of the prior models. The basic principle underlying boosting is to concentrate on misclassified cases and award them larger weights throughout the training phase, emphasizing the difficult-to-classify occurrences.

1. Assignment initial weights
2. Model training
3. Weight update
4. Sequential model training
5. Final prediction

Ques 75
AdaBoost and Gradient Boosting are both strong algorithms capable of producing very accurate models. Because it provides greater weights to misclassified samples, which might be impacted by outliers, AdaBoost is more sensitive to noisy data and outliers. Gradient Boosting, on the other hand, handles outliers and noisy data better due to its residual-based optimisation technique.

Ques 76
The goal of Random Forests in ensemble learning is to combine the predictions of numerous decision trees to construct a robust and accurate model. Random Forests are a popular ensemble approach that takes advantage of the notion of bagging (Bootstrap Aggregating) and adds a random aspect to the tree-building process.

Ques 77
Random Forests give a measure of feature relevance depending on how much each feature adds to the ensemble's prediction accuracy or quality. The significance of a

feature is measured by the drop in model performance when that feature is randomly permuted or deleted from the dataset.

1. Tree based feature importance
2. Aggregation
3. Normalization
4. Feature ranking

Ques 78

Stacking, also known as stacked generalization, is an ensemble learning approach that uses a meta-model to aggregate the predictions of numerous models. It employs a two-level learning process in which the base models predict based on the input data and the meta-model learns to predict based on the outputs of the base models.

1. Base model training
2. Creating intermediate predictions
3. Meta model training
4. Final prediction

Ques 79

Advantages

1. Improved predictive performance
2. Increased stability
3. Robustness
4. Feature importance

Disadvantage

1. Increased complexity
2. Difficulty in interpretability
3. Potential overfitting
4. Computational requirements

Ques 80

There is no one-size-fits-all method for determining the best number of models to include in an ensemble. It frequently necessitates experimenting, taking into account the individual characteristics of the situation and available resources. Cross-validation, learning curve analysis, and computational considerations, among other techniques,

can assist guide decision-making and help discover a reasonable balance between performance and resources.