**MACHINE LEARNING**

1. **R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?**

**Sol:-** R-squared and Residual Sum of Squares (RSS) are both Measures used to assess the goodness of fit of a regression model, but they capture different aspects of model performance.

R-squared, also known as the coefficient of determination, represents the proportion of the variance in the dependent variable that is predictable from the independent variables in the model. It ranges from 0 to 1, with 1 indicating a perfect fit. R-squared is a relative measure. As it compares the performance of the fitted model to a simple average of the dependent variable. One of the advantages of R-squared is the dependent variable variation that is explained by the independent variable variation that is explained by the independent variable.

On the other hand, the Residual Sum of Squared(RSS) Measures the total variance of the data that is not explained by the model. It calculates the sum of the squared differences between the observed values and the values predicted by the model. Therefore, a smaller RSS indicates a better fit of the model to the data. R-saured is often more commonly used and easier to interpret, as it provides a clear indication of the proportion of variance explained by the model.

**2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.**

1. **Sol:-** Total sum of squares (TSS) :- TSS measures the total variability in the dependent variable (y).It represented the sum of the squared difference between each observed values of the mean, regardless of the model. Mathematically, TSS is calculated as:

TSS = TSS = Σ(yᵢ - ȳ)².

1. Explained sum of Squares (ESS) :- ESS measures the variability in the dependent variable that is explained by the regression model. It represents the sum of the squared differences b/w the predicted values of the dependent variable and the overall mean of the dependent variable . Mathematically, ESS is calculated as:

ESS = ESS = Σ(ŷᵢ - ȳ)²

1. Residual sum of Squared (RSS):- RSS measures the variability in the dependent variable that is not explained by the regression model. It represents the sum of the squared differences b/w the observed values of the dependent variable and the predicted values

from the regression model.

RSS = Σ(yᵢ - ŷᵢ)²

The relationship b/w these metrics can be expressed using the following equation: TSS = ESS + RSS

**3. What is the need of regularization in machine learning?**

**Sol:-** Regularization in machine learning is employed to address the issue of overfitting, which occurs when a model performs well on the training data but fails to generalize to new,unseen data. The need for regularization arises from the following key reasons:

1. Overfitting.
2. Feature Selection.
3. Improved Generalization.
4. Mitigating Multicollinearity.
5. Bias-Variance Tradeoff.

The need for regularization in machine learning is driven by its ability to prevent overfitting, improve generalization performance, handle mulicolinearity,facilitate feature selection, and manage the bias-variance tradeoff,ultimately leading to more robust and realiable predictive models.

1. **What is Gini–impurity index?**

**Soln:-** The Gini impurity index is a measure used in decision trees and random forests to evaluate the impurity or homogeneity of set of data points. It’s used as a criterion for deciding the optimal split at each node of a decision tree and plays a crucial role in determining the feature that best separates the data.

The Gini impurity index is a matrix is a metric used in decision tree algorithm to evaluate the impurity or purity of set of data points based on their class labels. It’s a key component in determining the optimal feature and split points for building an effective decision tree model.

1. **Are unregularized decision-trees prone to overfitting? If yes, why?**

**Soln:-** Yes, unregualarized decision trees are prone to overfitting.Overfitting,Occurs when a model learns the training data too well,capturing noise and spurious patterns that do not generalize to new,unseen data.Decision trees, in their unregularized form, are susceptible to overfitting due to several reasons:

1.High variance.

2.Memorization of training data.

3.Lack of Global Control.

4.Insensitivity to small Changes.

5.Limited Generalization.

**6.What is an ensemble technique in machine learning?**

**Soln:-** Ensemble techniques in machine learning involve the combination of multiple models to create a more robust and accurate predictive model than any individual model alone. The main idea behind ensemble methods is to leverage the diversity of multiple models to improve predications,capture different aspects of the data,and mitigate the weakness of individual models. There are several types of ensemble techniques, some of the most popular ones include:

1.Bagging (Bootstrap Aggregating)

2.Boosting

3.Stacking(Stacking Generalization)

4.Voting

Ensemble techniques are effective because they can reduce overfitting, improve generalization, and produce more stable and accurate predictions by leveraging the strength of diverse models.By combining the predictions of multiple models,ensembles often outperform individual models,making them a popular and powerful tool in machine learning.

**7. What is the difference between Bagging and Boosting techniques?**

**Soln:-** The Bagging and Boosting techniques are both ensemble learning methods used in machine learning, but they differ in their approach to building models and combining their predictions.

The main difference b/w bagging and boosting lie in their training approaches- bagging trains models independently and then combines their predictions, while boosting builds models sequentially by altering the weights of training examples to focus on difficult instance.Despite their difference , both technique are powerful tools in ensemble learning ,aiming to improve predictive accuracy and generalization performance.

**8. What is out-of-bag error in random forests?**

**Soln:-**  In the context of random forest, the out-of-bag (OOB) error is an error estimation method that leverages the bootstrap aggregating(bagging) technique used in constructing the random forest model.When building each decision tree in random forest model.when building each decision tree in random forest models.when building each decision tree in random forest, a subset of the original training data is samoled with replacement,resulting in some data points being left out of the training set for that particular tree.These left-out data points from the out-of-bag samples for each tree.

The out-of-bag error in random forests offers an efficient and practical way to estimate the generalization error of the model without requiring a seprate validation set, thereby providing a valuable tool for assessing the models predictive performance.

**9. What is K-fold cross-validation?**

**Soln:-** K-fold cross-validation is popular technique used to assess the performance of a machine learning model and to ensure that the model’s performance is consistent across different subsets of the data. The process involves dividing the dataset into k equal-sized folds, where one fold is used as the validation set and the remaining k-1 folds are used exactly once as the validation data. The k results

Can then be averaged to produce a single estimation.

The steps for K-folds cross-validation are as follows:

1.The dataset is divided into k subsets of approximately equal size.

2. The model is trained k times each time using a different folds as the validation set and the remaining k-1 folds as the training set.

3.The performance metric(e.g., accuracy, error rate) is calculated for each validation set.

4.The k performance metrics are averaged to obtain a single performance estimation ,which is used to assess the generalization performance fo the model.

**10. What is hyper parameter tuning in machine learning and why it is done?**

**Soln:-** Hyperparameter tuning,also known as hyperparameter optimization,refers to the process of selecting the optimal hyperparameter values for a machine learning model.Hyperparameter are configurations external to the model that control its learning process.They are not learned from the data but are set before the learning process begins.Some common hyperparameters include the learning rate,the number of hidden layers in neural network,the depth of a decision tree,and the kernel type in support vector machine.

Hyperparameter tuning is performed for the following reasons:

1.Model Performance.

2.Overfitting and Underfitting.

3. Generalization

4.Computational Efficiency.

Hyperparameter tuning is a critical step in the machine learning model development process ,aimed at selecting the most effective to maximize the model’s performance, reduce overfitting and underfitting , improve generalization , and enhance computational E ffiviency.

**11. What issues can occur if we have a large learning rate in Gradient Descent?**

**Soln;-** If the learning rate in Gradient Descent is set too high,sevral issues my arise,potentially hindering the convergence and optimization process of the model. Some of the key issues associated with a large learning rate include.

1. Divergence.

2. Instability.

3. Overshooting.

4. Suboptimal Solutions.

5. Long Training times.

**12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?**

**Soln;-** Logistic Regression is a linear model,which means it assumes a linear relationship between the features and the log-odds of the target variable.AS a result,when the decision boundry b/w classes in the data is non-linear, Logistic Regression may struggle to effectively model the relationship and accurately classify the data.

When the classes in a dataset are not linearly separable,meaning a straight line cannot effectively seprate the classes,logistic Regression’s inability to capture non-linear patterns can lead to poor classification performance. In such cases,the model may underfit the data, leading to suboptimal predictive accuracy.

**13. Differentiate between Adaboost and Gradient Boosting.**

**Soln:-** Adaboost (Adaptive Boosting) and Gradient Boosting are both ensemble learning methods that combine multiple weak learners to create a strong learner. However, they differ in their approach to building the ensemble model.

Adaboost:

1.Adaboost sequentially trains a series of weak learners, with each new model correcting the errors made by its predecessors.

2.It assigns higher weights to misclassified data points, thus forcing the subsequent weak learners to focus more on the misclassified instances.

3.The final prediction is made by combining the predictions of all weak learners through a weighted majority vote.

Gradient Boosting:

1.Gradient Boosting builds the ensemble model in a stage-wise fashion, where each new model fits the residual errors (the gradient) of the preceding model.

2.It uses gradient descent algorithm to minimize the loss when adding new models to the ensemble.

3.The final prediction is the sum of the predictions of all the weak learners, where each weak learner is assigned a weight based on its contribution to minimizing the overall loss function

**14. What is bias-variance trade off in machine learning?**

**Soln:-** The bias-variance trade-off is a fundamental concept in machine learning that describe the blance b/w the error introduced by the bias of the model and the error introduced by the variance of the model.

1.Bias:- bia referce to the error introduced by approximating a real-word problem with problem with a simplified model. High bias can lead to underfitting, where the model is unable to capture the complexities of the data,reslting in poor performance on both the training and test datasets.

2.Variance: Variance, on the other hand,refers to the model’s sensitivity to the variability in the training data.High variance can lead to overfitting, where the model performs well on the training data but poorly on unseen test due to capturing noise and not generalizing well.

The bias-variance trade-off highlight the delicate balance b/w model simplicity and flexibility, and finding this balance is crucial for developing machine learning models that generalize well to new unseen data.

**15. Give short description each of Linear, RBF, Polynomial kernels used in SVM**

**Soln:-**  1.Linear Kernel

* The linear kernel is the simplest kernel function used in SVM.
* It computes the inner product of the input features and is suitable when the data is linearly separable.

2.RBF(Radial Basis Function) Kernal:

* The RBF kernel is popular choice in SVM for handling non-linearly seprabledata.
* It measures the similarity b/w two sample based on the distance b/w them in the feature space.

3.Polynomial kernel:

* The polynomial kernel is used to handle non-linear data through transformation into a higher – dimensional space.
* It computes the similarity b/w samples as the power of the sum of the original features, making it effective for capturing non-linear relationships.