MACHINE LEARNING (Answers)

1. R-squared is a better measure of goodness of fit model in regression.
   Reason: RSS is sensitive to the scale of the dependent variable. If we change the units of the dependent variable, the RSS value will change. This makes it difficult to compare models with different scales. Other than this, A model with more parameters can achieve a lower RSS by fitting noise or outliers. So overfitting issue is present in RSS. Outlier data points can affect the RSS value significantly.

2. TSS (Total Sum of Squares) : Measurement of TSS is given by

$$\text{TSS} = \sum_{i=0}^{n} (yi - y)^2$$

Where yi=Actual value and y= mean value

RSS (Residual Sum of Squares) : Measurement of RSS is given by

$$\text{RSS} = \sum_{i=0}^{n} (yi - y*)^2$$

Where yi=Actual value and y*= predicted value

3. When we use regression model to train some data there is good chance that the model will overfit the given training dataset. So regularization helps to sort this overfitting problem by restricting the degree of freedom of a given equation. In a linear equation we don't want huge coefficients as a small change in coefficient can make a large difference for the response variable. So basically to avoid overfitting regularization is required.

4. The Gini Impurity index helps in identifying the most suitable feature for node splitting while construction of a decision tree classifier. it is a method that measures the impurity of a dataset. The more impure the dataset, the higher is the Gini index. The word “Impurity” indicates the number of classes present in a subset. The more distinct classes included in a subset then higher the impurity present.

5. Yes, unregularized decision-trees prone to overfitting because decision trees can overfit when there is limited training data. Other than this when decision trees grow more deeper, there will be more chances for overfitting.

6. This topic is not covered by Datatrained till now.
7. This topic is not covered by Datatrained till now.
8. This topic is not covered by Datatrained till now.

9. K-fold cross-validation is used for evaluating models. When we split data into training and test data set, then train the model with the training set and evaluate the result with test set. So we have evaluated the model only one time and we are not sure that good accuracy of model is real or not. So we want to evaluate the model multiple times to cross check the model accuracy. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

10. Hyperparameter tuning is the process of selecting the optimal value for model’s hyperparameters. Hyperparameters are used to control the behaviour of model by selecting the best the learning rate of the model.

11. If we have a large learning rate in gradient descent there will be more chances that the model will not trained at its optimal way. The large learning rates result in quick changes and require fewer training. A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution.

12. No, we can’t use Logistic Regression for classification of Non-Linear Data because logistic regression works on the formula (y=mx+c where y=label , x=data, m=coefficient/slope & c=intercept) which means it only works for linearly separated data.

13. This topic is not covered by Datatrained till now.

14. When we try to decrease the value of variance then bias will increase and when we try to decrease the value of bias then variance will increase. So we have to settle the value of bias and variance at some point in between. It is called bias variance trade off.

15. This topic is not covered by Datatrained till now.