Daily Machine Learning Interview Questions

41. When should Classification be used over Regression?

Both classification and regression are associated with prediction. Classification involves the identification of values or entities that lie in a specific group. Regression entails predicting a response value from consecutive sets of outcomes

Classification is chosen over regression when the output of the model needs to yield the belongingness of data points in a dataset to a particular category

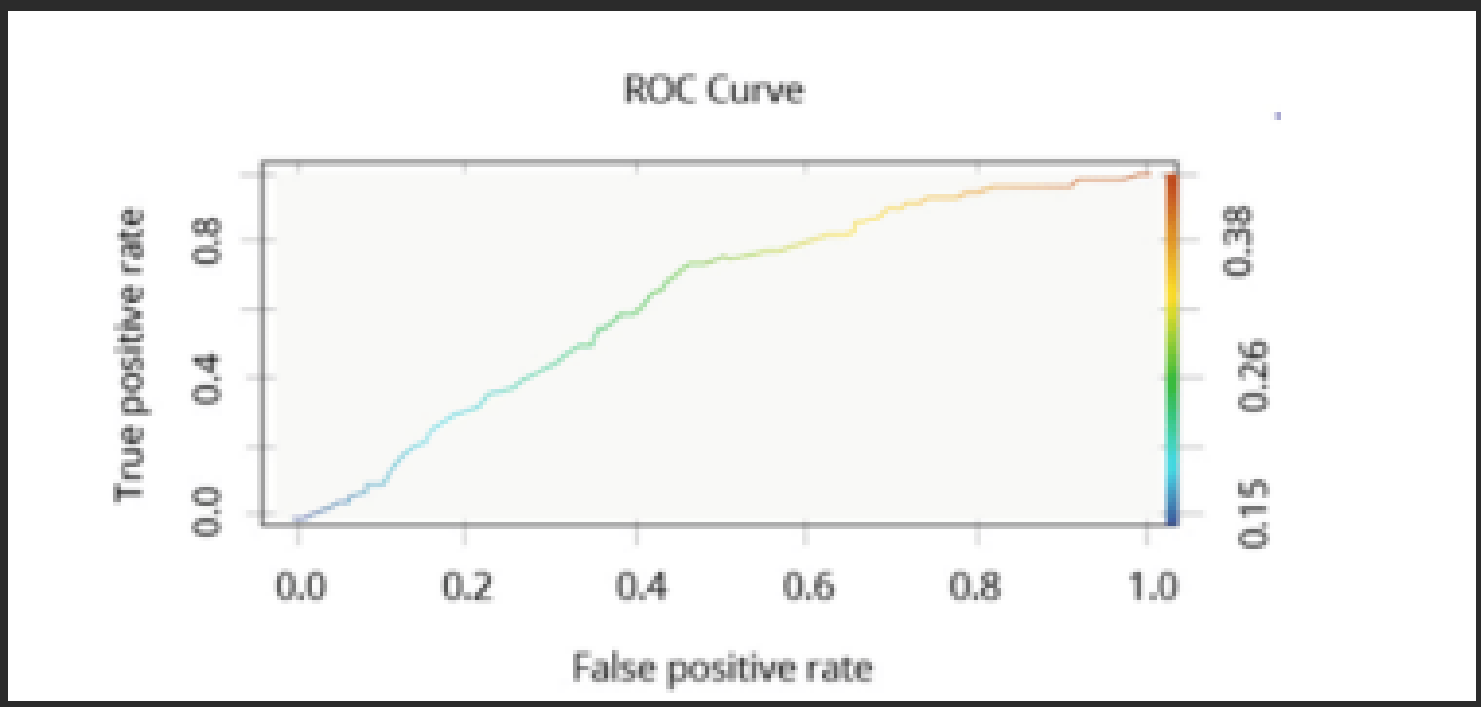
For example:

If you want to predict the price of a house, you should use regression since it is a numerical variable. However, if you are trying to predict whether a house situated in a particular area is going to be high-, medium-, or low-priced, then a classification model should be used.

42. What is ROC Curve and what does it represent?

ROC stands for receiver operating characteristic.

ROC Curve is used to graphically represent the trade-off between true and false positive rates. In ROC, area under the curve (AUC) gives an idea about the accuracy of the model.



The above graph shows an ROC curve. The greater the AUC, the better the performance of the model

43. Why are Validation and Test Datasets Needed?

Data is split into three different categories while creating a model:

- Training dataset: Training dataset is used for building a model and adjusting its variables. The correctness of the model built on the training dataset cannot be relied on as the model might give incorrect outputs after being fed new inputs.

- Validation dataset: Validation dataset is used to look into a model's response. After this, the hyperparameters on the basis of the estimated benchmark of the validation dataset data are tuned. When a model's response is evaluated by using the validation dataset, the model is indirectly trained with the validation set. This may lead to the overfitting of the model to specific data. So, this model will not be strong enough to give the desired response to real-world data

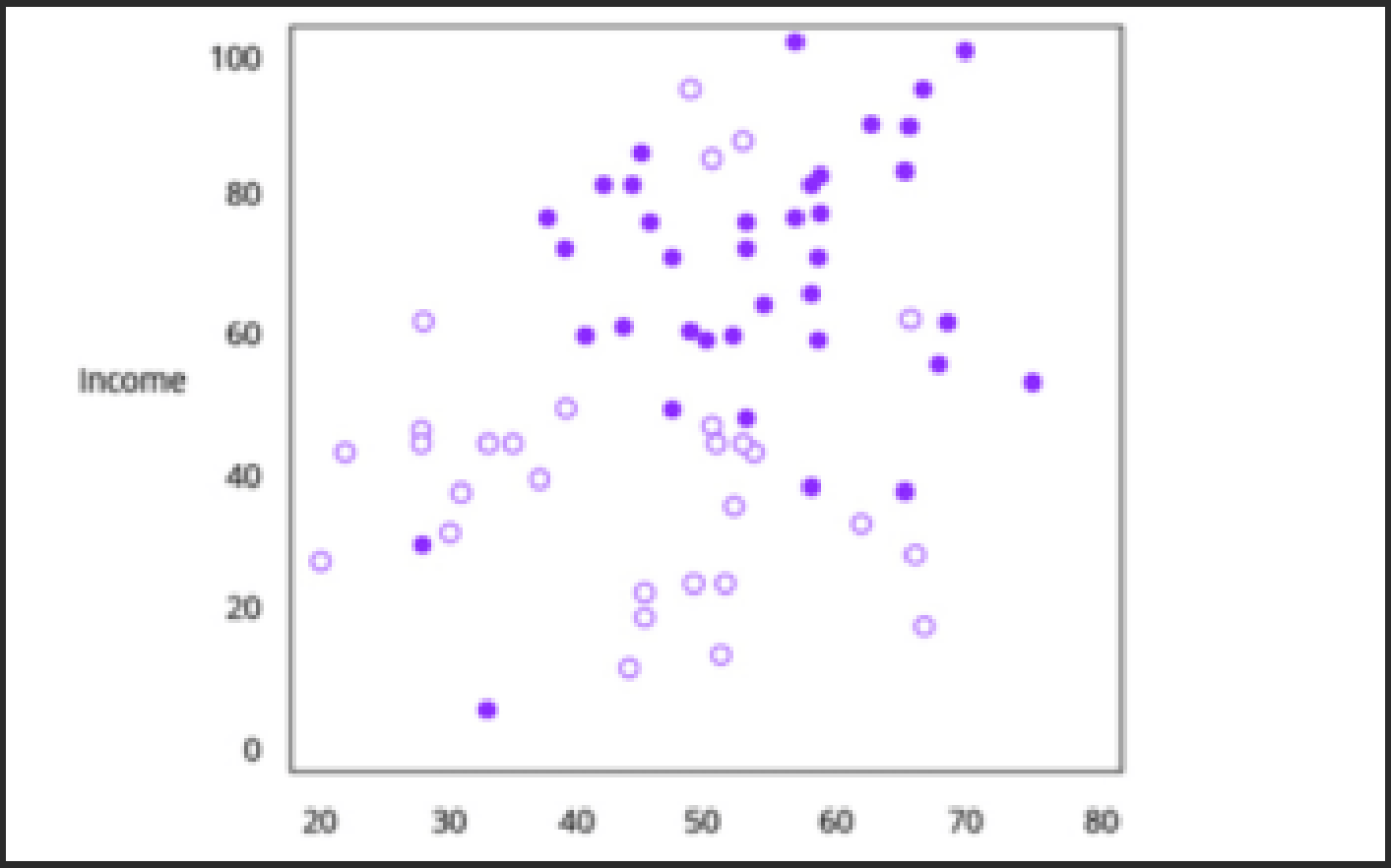
- **Test dataset: Test dataset is the subset of the actual dataset, which is not yet used to train the model. The model is unaware of this dataset. So, by using the test dataset, the response of the created model can be computed on hidden data. The model's performance is tested on the basis of the test dataset.**

Note: The model is always exposed to the test dataset after tuning the hyperparameters on top of the validation dataset.

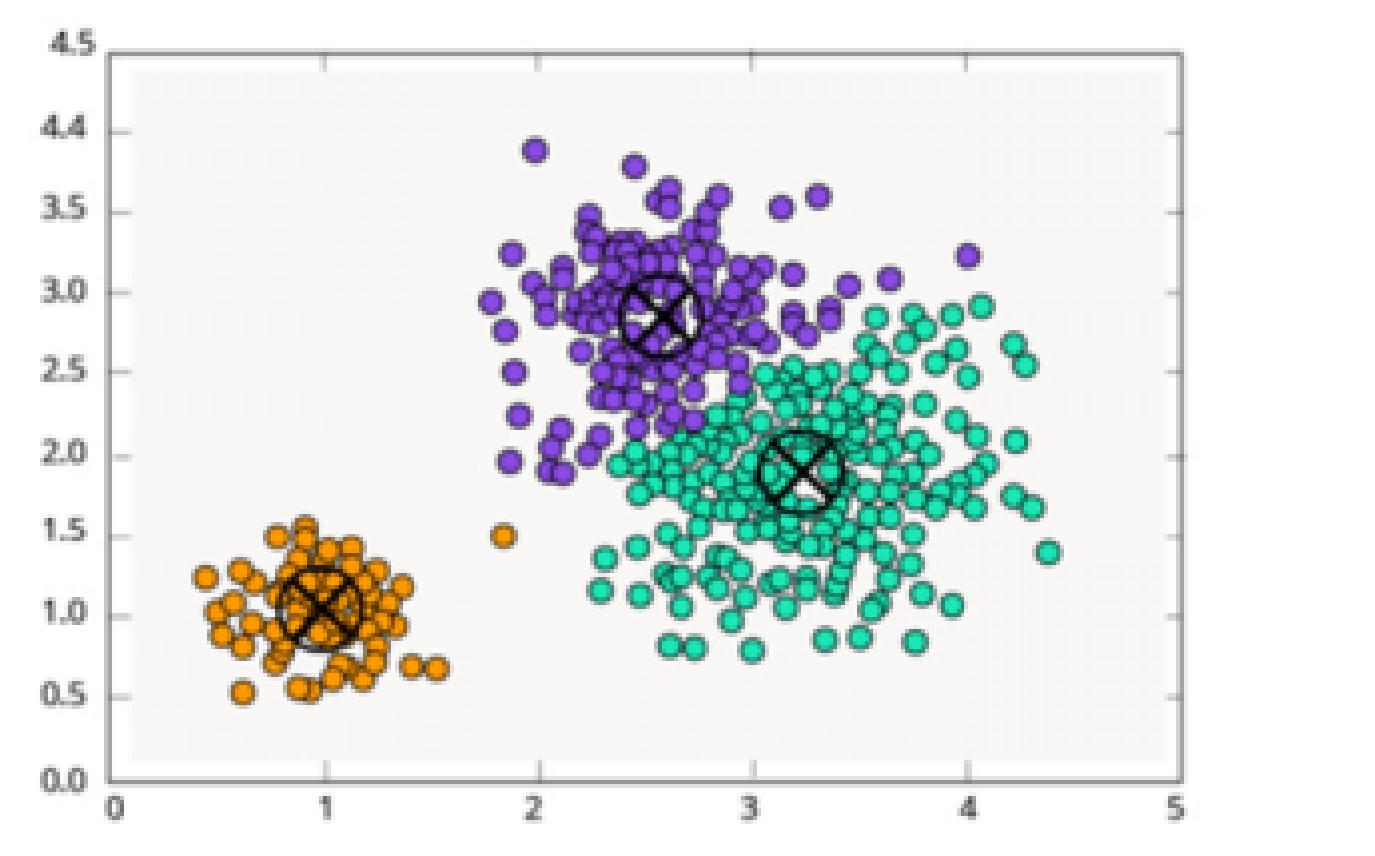
As we know, the evaluation of the model on the basis of the validation dataset would not be enough. Thus, the test dataset is used for computing the efficiency of the model

44. Explain the difference between KNN and K-means Clustering

- K-nearest neighbors (KNN): It is a supervised Machine Learning algorithm. In KNN, identified or labeled data is given to the model. The model then matches the points based on the distance from the closest points.



- K-means clustering: It is an unsupervised Machine Learning algorithm. In K-means clustering, unidentified or unlabeled data is given to the model. The algorithm then creates batches of points based on the average of the distances between distinct point



45. What is Dimensionality Reduction?

In the real world, Machine Learning models are built on top of features and parameters. These features can be multidimensional and large in number. Sometimes, the features may be irrelevant and it becomes a difficult task to visualize them.

This is where dimensionality reduction is used to cut down irrelevant and redundant features with the help of principal variables. These principal variables conserve the features, and are a subgroup, of the parent variables

Thank You