1. Linear Regression: Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple. The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error are as small as possible. Error is the distance between the point to the regression line.

2: Ridge and Lasso Regression:

Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression.

3. KNN : The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It’s easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (classification) or averages the labels (regression).

5. Cross Validation: Similar to train/test split, but it’s applied to more subsets. We split our data into k subsets, and train on k-1 one of those subsets. We test the model on the last subset. We can do this for all the subsets.

6. Train test split: The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data. We have the test dataset in order to test our model’s prediction on this subset. It helps to determine accuracy.

7. Loss function:  It’s a method of evaluating how well a specific algorithm models the given data. If predictions deviate too much from actual results, loss function would be a very large number. It tell us how much error is there in the model.

8. Generating seaborn heatmap:  Seaborn library is built on top of Matplotlib. Seaborn library provides a high-level data visualization interface. A Heatmap is a way of representing the data in a 2-dimensional form. The data values are represented as colors in the graph. The goal of the heatmap is to provide a colored visual summary of information.

9. Reshaping: It is a function under numpy. The reshape() function is used to give a new shape to an array without changing its data. Changes the dimensions o the array basically.

10.

Underfitting: When a model is underfitted, it means that the model does not fit the training data and therefore misses the trends in the data. The model cannot be generalized to new data.

Overfitting: The model we have trained is fit too closely to the training set.  This model will be very accurate on the training data but will not be very accurate on untrained or new data.

12. Confusion matrix: A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.  It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.4

13. Classification Report: A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False.The report shows the main classification metrics precision, recall and f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives.

14. Logistic regression: It is basically a supervised classification algorithm. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

15. ROC curve :

Sensitivity measures the proportion of positives that are correctly identified as such.

Specificity measures the proportion of negatives that are correctly identified as such.

Receiver Operating Characteristic Curve can be obtained by plotting sensitivity on y-axis and specificity on x-axis. The further the curve is from the diagonal line, the better the model is at discriminating between positives and negatives in general.

16. AUC: AUC stands for Area under the curve. AUC gives the rate of successful classification by the logistic model. The AUC makes it easy to compare the ROC curve of one model to another. Greater the area better the model.

18. Pipeline: