**MACHINE LEARNING – WORKSHEET 6**

Solution:1 – C

Solution:2 – C

Solution:3 – C

Solution:4 – C

Solution:5 – A

Solution:6– D

Solution:7 – B

Solution:8 – A

Solution:9 – B

**Solution:10 –** Adjusted R-Squared is a modified version of R- Squared adjusted with the number of predictors. It penalizes for adding unnecessary features and allows a comparison of regression models with a different number of predictors.

R-Squared increases even when you add variables which are not related to the dependent variable, but adjusted R-Squared take care of that as it decreases whenever you add variables that are not related to the dependent variable, thus after taking care it is likely to decrease.

**Solution:11** – Lasso:- The LASSO method aims to produce a model that has high accuracy and only uses a subset of the original features. The way it does this is by putting in a constraint where the sum of the absolute values of the coefficients is less than a fixed value. To that end it lowers the size of the coefficients and leads to some features having a coefficient of 0, essentially dropping it from the model. In this way, it is also a form of filtering your features and you end up with a model that is simpler and more interpretable.

However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.

Ridge:- The Ridge Regression method was one of the most popular methods before the LASSO method came about. The idea is similar, but the process is a little different. The Ridge Regression also aims to lower the sizes of the coefficients to avoid over-fitting, but it does not drop any of the coefficients to zero. The constraint it uses is to have the sum of the squares of the coefficients below a fixed value. The Ridge Regression improves the efficiency, but the model is less interpretable due to the potentially high number of features.

**Solution:12** – The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

Values of VIF that exceed 10 are often regarded as indicating multicollinearity, but in weaker models values above 2.5 may be a cause for concern. In many statistics programs, the results are shown both as an individual R2 value (distinct from the overall R2 of the model) and a Variance Inflation Factor (VIF).

**Solution:13** – The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.

Suppose we have two features of weight and price, as in the below table. The “Weight” cannot have a meaningful comparison with the “Price.” So the assumption algorithm makes that since “Weight” > “Price,” thus “Weight,” is more important than “Price.”

So these more significant number starts playing a more decisive role while training the model. Thus feature scaling is needed to bring every feature in the same footing without any upfront importance. Interestingly, if we convert the weight to “Kg,” then “Price” becomes dominant.

Another reason why feature scaling is applied is that few algorithms like Neural network gradient descent converge much faster with feature scaling than without it.

**Solution:14** – A good regression model is one where the difference between the actual or observed values and predicted values for the selected model is small and unbiased for train, validation and test data sets.

To measure the performance of your regression model, some statistical metrics are used. Here we will discuss four of the most popular metrics. They are-

Mean Absolute Error(MAE)

Root Mean Square Error(RMSE)

Coefficient of determination or R2

Adjusted R2

**Solution 15:**

Sensitivity – 80

Specificity – 96

Precision – 0.952

Recall – 0.8

Accuracy – 0.88