***Assignment – 03***

***Spring 2021***

***Answer to the question number (1)***

PCA is considered only as a variance to the features, but it is not considered as a relationship between features and labels while doing feature reduction.

On other hand, regularization acts directly on the relationship between features and labels. Thus, regularization helps develop models that are better at showing predictions for a given set of features.

***Answer to the question number (2)***

F1 score is the balance between the precision and the recall.

F1= 2\*((precision\*recall)/(precision + recall))

F1 score used as a performance metric for classification algorithms and gives a better measure of the incorrectly classified cases than the accuracy metric.

The methods we can use to handle missing or corrupted data are:

**Method 1 is deleting rows or columns:**  
We usually use this method when it comes to empty cells. If the majority of our data is missing for a column or for a row, we can simply delete them.

**Method 2 is replacing the missing data with aggregated values:**  
In this case, we can calculate the aggregated value based on the rest of the values we have in the column and put the received number to the empty spot.

**Method 3 is creating an unknown category:**  
Categorical features have a number of possible values, so we can create one more category for the missing values. This way we will lower the variance by adding new information to the data. This could be used when the original information is missing or cannot be understood,

**Method 4 is predicting missing values:**where we have no missing values, we can train a machine learning algorithm in order to predict the missing values. Since the samples for which this training is performed, there are missing values, it is necessary to replace them initially using one of the simplest methods for recovering gaps. This way will give us better performance, unless, of course, a missing value should have a high variance.

***Answer to the question number (3)***

*Principal component are the eigenvectors of a covariance matrix. The idea to reduce the dimensionality of the dataset by transforming the variables to a new set of variables, known as the principal components, ordered in such way the retention of variation present in the original variables decreases as we move down.*

*The difference between the first and the last principal component is that the first principal component has the direction in space along which projections have the largest variance, whereas the last component has the variance-maximizing direction orthogonal to the previous component.*

*KNN algorithms is said to work better on small data sets because when the dataset is small, the classifier completes execution in shorter time duration. The algorithm doesn’t have a training step because it is a case-based learning algorithm to make predictions.*

***Answer to the question number (4)***

*Kernel trick is used for making the machine learning model fit the training data well so that the model can learn the trends and the patterns present within the data so that it can classify the data points really well.*

*Thus, the kernel trick allows the model to operate in the original feature space without computing the coordinates of the data in a higher-dimensional space.*

*No, if we use this trick for arbitrary dimensions for resulting optimization, a significant computational burden might occur.*

***Answer to the question number (5)***

*When the max\_depth value is none, it is set in default mode and the nodes are derived from a single class, the decision tree takes its natural course of expanding and is not expanded forcefully. When the max\_depth value is higher than the default value, the nodes are derived from more than one class and it causes overfitting of the decision tree model. When the max\_depth is lower than the default value, the nodes are derived from a single class and it causes the underfitting of the decision tree model.*

*No, all leaf nodes will not have data from a single class. The max\_depth is used when you need the discrete value or more detailed value from a decision tree classifier. When you want to expand the decision tree more than its course. It is still preferred sometimes for tuning of the model.*

*A node is said to be impure if it is split from different classes for example if a node is 100% derived from a single class then it is pure but if a node is 50% derived from one class and 50% derived from another then the node is impure.*

*Impurity can be minimized by keeping the value of max\_depth minimum but it may cause underfitting of the model.*

***Answer to the question number (6)***

*There is a threshold to the number of trees that should be used, beyond which there is no significant performance gain, and would only increase the computational cost.*

*A Random Forest is a supervised learning algorithm that performs both regression and classification tasks with the use of multiple decision trees. The "forest" it builds, is an ensemble of decision trees and this technique is known as bagging. The idea of the bagging method is a combination of learning models increases the overall result. It can be used for improving both classification and regression problems.*