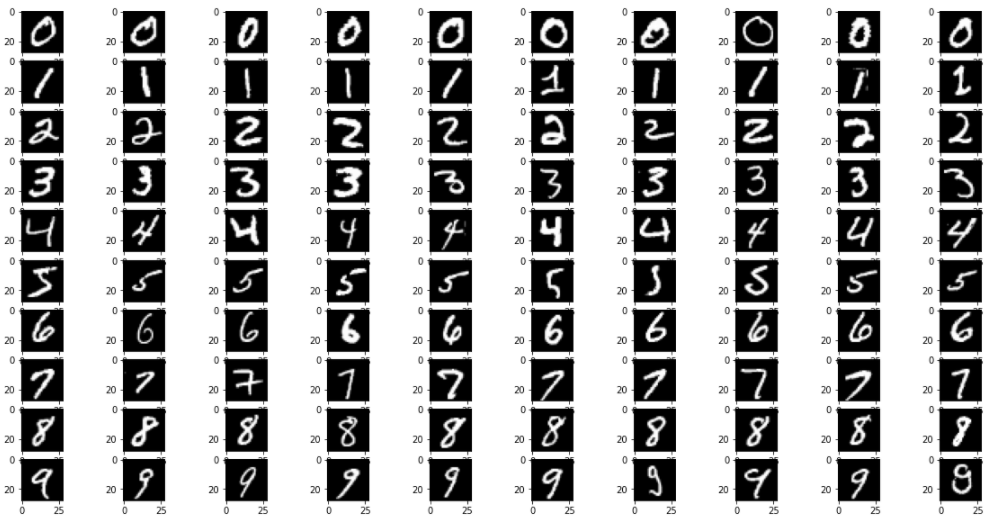
**Analysis Report**

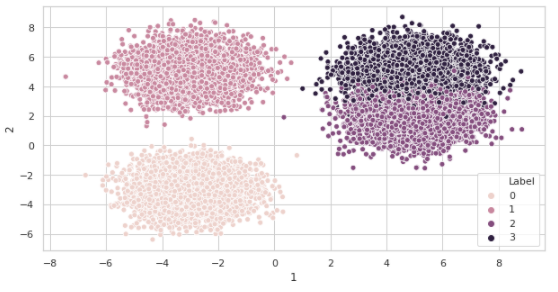
**1(a):**



**Observation:**

1. The dataset\_1 contains the samples of images where each image belong to the one of the Digit(0 - 9).

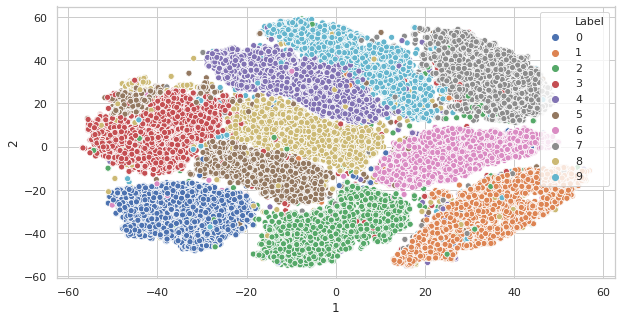
**1(b):**



**Inference:**

1. The dataset consists of data points belonging to one of the four classes, i.e., 0, 1, 2, and 3.
2. The data points with classes 1 and 0 are well separated, as inferred from the figure.
3. The data points with classes 2 and 3 are not well separated. This overlapping is because of some similarities in the shape of 2 and 3.

**1(c):**

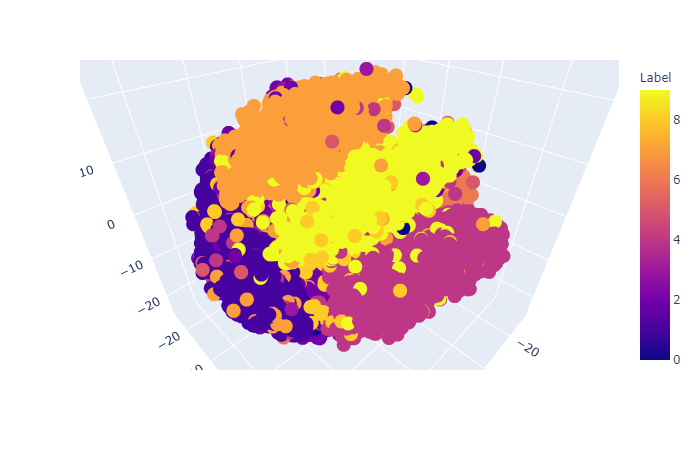
****

**Inference:**

After reducing the data in the 2Dimesnional space, the visualization of different classes appears as the clusters where all data points in the same clusters belong to one of the possible classes.

1. The most well-separated classes in this dataset are 0, 6, and 1.
2. The least separation for the data points happens for classes 5, 3, and 8.

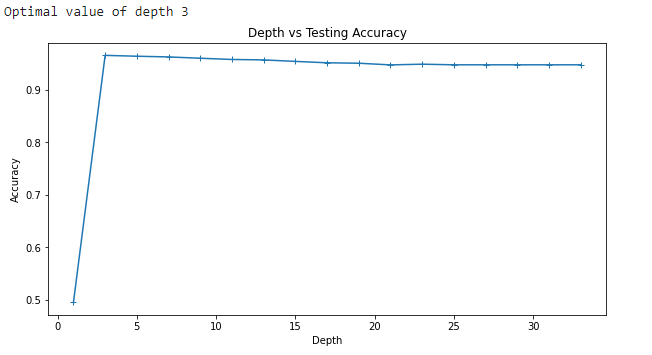
**1(d):**



**Distinct with the previous plot :**

1. In this plot, the separation for the classes is decreased by some factor.
2. In the 2D scatter plot, the separation of classes appears to form the plot itself, while, in this case, the plot appears to be more overlapping.

**2(a):**



**Effect of Depth on the Model :**

1. Initially, the depth of the Model is very low, both training and testing accuracy is low, which means that model under fits.
2. As the depth reaches 3, the testing accuracy is high, showing the model fits the data.
3. As the depth increases, the model’s testing accuracy starts decreasing, which shows that model overfits.

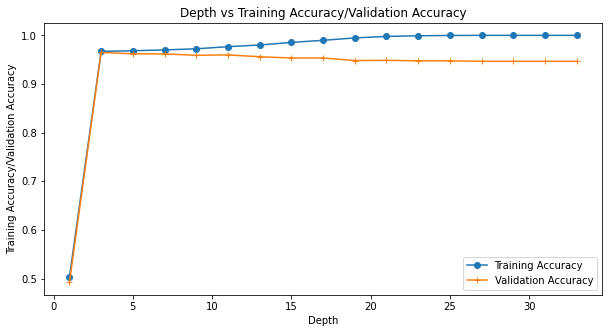
**The best performance of the model is consistent with the visualization in 1(b) because :**

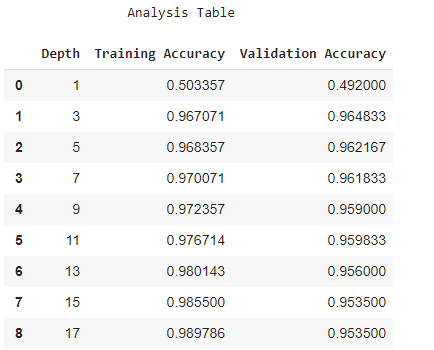
1. The visualization shows that class separation happens for the two classes, while the other two classes have overlapping data instances.
2. When the depth is taken to be 3, then such class separation is possible with the depth 3, with most separated classes appearing in with depth 1 and depth 2 while the overlapping classes in depth.
3. As the depth increase form 3, the model tries to lear the instances that belong to the overlapping classes, which causes the model to overfit.

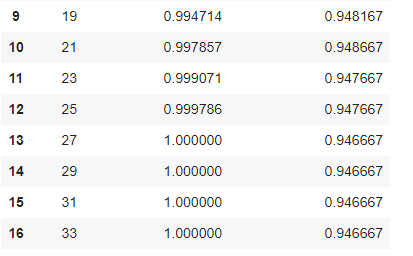
**Splitting the DataSet :**

1. For splitting the dataset into training and testing, a sample function forms the pandas libraries is used that extracts the given number of samples randomly.

**2(b):**

****





**Depth 1:** Underfitting because both training and validation accuracy is low.

**Depth 3:**  Neither Underfitting nor overfitting because both training and validation accuracy is high.

**Depth 5:** The model overfits because the training accuracy has increased as compared to previous training accuracy while the validation accuracy has decreased as compared to previous validation accuracy.

**Depth 7:** The model overfits because the training accuracy is increasing while the validation accuracy is decreasing.

**Depth 9:** The model overfits.

**Depth 11:** The model overfits.

**Depth 13:** The model overfits.

**Depth 15:** The model overfits.

**Depth 17:** The model overfits.

**Depth 19:** The model overfits.

**Depth 21:** The model overfits.

**Depth 23:** The model overfits.

**Depth 25:** The model overfits

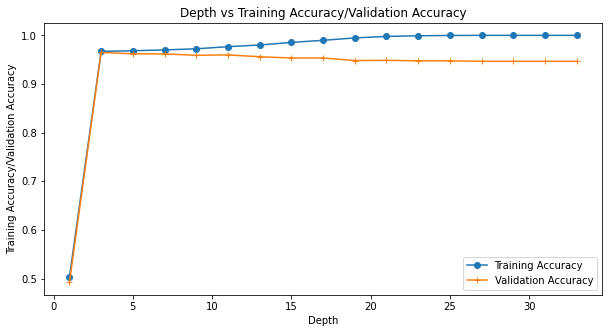
**Depth 27:** The model overfits.

**Depth 29:** The model overfits.

**Depth 31:** The model overfits.

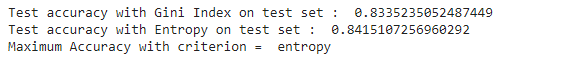
**Depth 33:** The model overfits.

**2(c):**

****

There is no deviation in the results produced by the implemented accuracy function and the library function used to calculate accuracy.

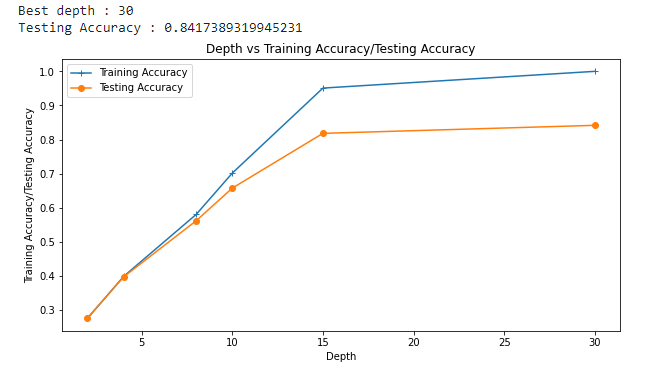
**3(a):**



**Data Preprocessing :**

1. To drop the column NO, we use the drop function of the panda's library.
2. To deal with the missing values in the column PM2.5, filling(method =bill) is used.
3. This will fill the missing values for the PM2.5 from the backward direction, i.e., it fills the null value form the next filled value in that column.

**3(b):**



Form the plot; it can be seen that training accuracy and testing are increasing and highest for the depth = 30.

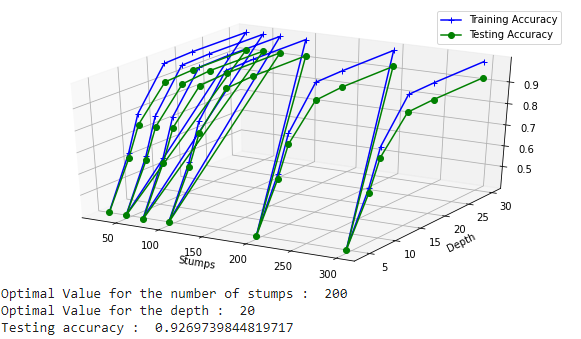
**3(c):**



**Inference :**

1. In comparison to the model in question 3(a) and 3(b), the accuracy has decreased drastically form 84 %(approx) to 34%(approx).
2. The most appropriate reason for this will be the underfitting of the model.
3. Although we are used the 100 decision trees to get better accuracy, each decision tree is highly biased because of the maximum depth of 3.
4. As a result, the accuracy is very low both for training and testing.

**3(d):**



|  |  |  |
| --- | --- | --- |
| **Model** | **Rank** | **Test Set Accuracy** |
| Default Model | 2 | 0.8481287083523505 |
| Depth: 30  Criterion: entropy | 3 | 0.8417389319945231 |
| No. of stumps: 100  Depth: 3  Criterion: entropy  Rule: Maximum Voting Rule | 4 | 0.3450479233226837 |
| No. of stumps: 200  Depth: 20  Criterion: entropy  Rule: Maximum Voting Rule | 1 | 0.9269739844819717 |