# DATA ANALYSIS - REPORT

## Supervised Lerning: Implementing Random Forest - RESULTS

|  |  |  |  |
| --- | --- | --- | --- |
| **B** | **Height** | **Training MSE** | **Test MSE** |
| 100 | 3 | 11.405249299711397 | 16.983355313717812 |

### **Understanding the effect of height**

|  |  |  |  |
| --- | --- | --- | --- |
| **B** | **Height** | **Training MSE**  **(Accurate to 2 decimal places)** | **Test MSE**  **(Accurate to 2 decimal places)** |
| 100 | 1 | 37.70 | 34.82 |
| 2 | 19.30 | 22.34 |
| 3 | 11.72 | 16.88 |
| 4 | 7.60 | 13.44 |
| 5 | 5.31 | 12.40 |
| 6 | 4.06 | 12.13 |
| 8 | 2.89 | 11.03 |
| 10 | 2.51 | 11.94 |
| 50 | 2.47 | 11.83 |

Chart, line chart

Description automatically generatedChart

Description automatically generated

Height or Tree Depth

Height or Tree Depth

**Observations**:

* Initially, test MSE is slightly better than training MSE when height of the tree is less than 2.
* Once we start increasing the height of the tree from 2 to 6, there is a sharp decline in both the MSEs with training MSE always lower than test MSE.
* Increasing the tree height further till 10 reduces training MSE but there is hardly any effect on test MSE.
* Any further increase in height brings negligible change in either of the MSEs.

**Discussion**:

* Trees very low in depth do not consider many features i.e., do not capture the data well so their training MSE is very low.
* Trees of more depth helps in achieving better accuracy as they are trained well with the training data.
* A tree depth capturing almost all features performs as good as trees with higher depth. Or, the improvement in accuracy due to increased height is negligible after a certain height where a significantly lower training MSE is achieved.

**Conclusion:**

To achieve a trade-off between computational cost and efficiency, a tree should not be very shallow, should be deep enough to consider significant amount of features at each node but should not be too deep as that increases the computational cost and time bringing negligible improvement in accuracy.

### **Understanding the effect of B**

|  |  |  |  |
| --- | --- | --- | --- |
| **B** | **Height** | **Training MSE**  **(Accurate to 2 decimal places)** | **Test MSE**  **(Accurate to 2 decimal places)** |
| 1 | 3 | 28.53 | 42.79 |
| 5 | 13.76 | 23.02 |
| 10 | 12.07 | 16.47 |
| 20 | 12.93 | 18.10 |
| 50 | 11.40 | 17.35 |
| 75 | 10.85 | 15.90 |
| 100 | 11.42 | 16.07 |
| 250 | 11.02 | 16.12 |
| 300 | 11.05 | 16.71 |
| 500 | 11.44 | 16.89 |
| 1000 | 11.12 | 16.28 |

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Number of trees

Number of trees

**Observations**:

* Training MSE is always lower than test MSE.
* For a very low value of B, test MSE is significantly higher than train MSE.
* Increasing the number of trees or B value helps in achieving better accuracy.
* After a certain value of B, there is negligible improvement in accuracy.

**Discussion**:

* Increasing the number of trees helps in lowering the variance and hence helps in achieving better test accuracy. [test error = irreducible error + bias2 + variance]
* Using only 1 tree for making prediction is not wise as it does not take advantage of bagging and ensemble method which help in achieving better accuracy.
* After a certain number of trees (large enough), there is negligible improvement in training and test MSEs. Hence, there is no point in increasing computational cost by taking very high value of B.

**Conclusion:**

To achieve a trade-off between computational cost and efficiency, an optimal number of trees should be considered (not very high or very low) where the number of trees is large enough to give an averaged predictions which are closer to the true values.

## Unsupervised Learning: Implementing Agglomerative Hierarchical Clustering – RESULTS

|  |  |
| --- | --- |
| **Linkage Function** | **Height of the root of Dendrogram** |
| Complete Linkage | 138.15044875568614 |
| Single Linkage | 93.06565171073733 |
| Average Linkage | 103.1596001630988 |
| Centroid Linkage | 84.53235880624236 |

### **Evaluating the Value of Potential Function**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **k** | **Value of Potential Function (accurate to 2 decimal places)** | | | |
| **Complete Linkage** | **Single Linkage** | **Average Linkage** | **Centroid Linkage** |
| 3 | 3805.97 | 3942.72 | 3859.08 | 3898.38 |
| 5 | 3615.26 | 3667.51 | 3685.37 | 3739.96 |
| 10 | 3118.20 | 3297.80 | 3105.97 | 3281.24 |
| 15 | 2770.95 | 2845.50 | 2741.79 | 3137.59 |
| 20 | 2491.65 | 2518.16 | 2392.41 | 5302.22 |
| 32 | 1715.65 | 1648.11 | 1669.65 | 3669.13 |
| 50 | 689.86 | 667.17 | 689.49 | 4156.91 |
| 60 | 151.71 | 151.71 | 151.71 | 231.85 |

Chart, line chart

Description automatically generated

**Observations:**

* The potential function value obtained for Complete and Average linkage is almost the same for any value of k i.e., number of clusters obtained after cutting the tree.
* The potential function value obtained for Single linkage is equally good for higher value of k but slightly high for lower values of k.
* Centroid Linkage gives higher values of potential function for higher value of k but performs equally well for lower values of k.

**Discussion:**

* *A lower potential function value indicates how close the observations are in every cluster for a cluster assignment which helps in better understanding the structure of the data.*
* For a cluster assignment obtained at lower value of k, we have less clusters and every cluster have observations which are closer to each other. This is true for all linkages as number of clusters is low to perceive an actual difference. So, we can use any of these linkage function to understand relation between observations when we have lower number of categories that these observations belong to (though Complete and Average perform slightly better).
* For a cluster assignment obtained at higher value of k, we have large number of clusters. These clusters have closer observations for complete, single, and average linkages. So, we can say that these linkages are able to better understand relation between observations when we have higher number of categories that these observations belong to.

**Conclusion:**

To understand the structure of the data, we should prefer using complete and average linkages as their cluster assignment gives cluster with closer observations for any number of clusters. The cluster assignment of single linkage also performs equally good except for smaller number of clusters. We must use centroid linkage only in case when number of clusters required is very low otherwise it performs poor cluster assignment.

### **Evaluating the maximum and minimum size of clusters formed in 2b)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **k** | **Complete Linkage** | | **Single Linkage** | | **Average Linkage** | | **Centroid Linkage** | |
| Min cluster size | Max cluster size | Min cluster size | Max cluster size | Min cluster size | Max cluster size | Min cluster size | Max cluster size |
| 3 | 3 | 42 | 1 | 62 | 2 | 54 | 1 | 61 |
| 5 | 2 | 40 | 1 | 57 | 1 | 54 | 1 | 59 |
| 10 | 1 | 21 | 1 | 52 | 1 | 31 | 1 | 52 |
| 15 | 1 | 19 | 1 | 39 | 1 | 20 | 1 | 49 |
| 20 | 1 | 9 | 1 | 34 | 1 | 16 | 1 | 46 |
| 32 | 1 | 6 | 1 | 19 | 1 | 8 | 1 | 34 |
| 50 | 1 | 3 | 1 | 4 | 1 | 3 | 1 | 3 |
| 60 | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 |

Chart, line chart

Description automatically generated

**Observations:**

* For lower values of k, the maximum size of cluster (out of k clusters obtained) is lower for complete and average linkage as compared to single and centroid linkage.
* For higher values of k, the maximum size of cluster converges for all the linkages.
* Centroid linkage overall generates largest size of cluster for every value of k followed by single linkage.

**Discussion:**

* For lower values of k, we are cutting tree at a height which is nearer to the root than the base. At this height, a very large maximum size of cluster as against to minimum size ranging between 1 to 3, points to the fact that the tree is more unbalanced. So, we can say that complete and average linkages generate more balanced dendrogram than single and centroid linkages.
* For higher values of k, we are cutting the tree at a height which is nearer to base, so the maximum size of cluster obtained is very low (as we have more chances of taking leaf nodes). All the linkages give the same size of maximum cluster with higher value of k.

**Conclusion:**

Complete and average linkages generate more balanced dendrogram as compared to single and centroid linkage which produce trailing clusters.

*End of Report*