**University Ranking**

**Data Cleaning**

The timesData dataset required the following data cleaning procedures:

1. Filtering the dataset for 2016 year.
2. Dropping the world\_rank column for fitting decision tree and other ensemble learning models using total\_score as response variable.
3. Removing rows with missing and NaN values.
4. Removing rows with hyphen values from income feature column.
5. Removing the commas in integer values in thousands separator in num\_students column and hence, converting from string to float.
6. Replacing the hyphens in total\_score column with a float value of 45 since all the values in this column below 48.8 were missing and denoted by hyphens.
7. Removing the % sign from international\_students column and hence, converting from string to float.
8. Removing rows with hyphen values in female\_male\_ratio column.
9. Converting the female\_male\_ratio column from ratio (string) to a decimal value (float).
10. Dropping the university\_name and country column from model prediction because all the rows have unique feature values and no two rows have the same values for which the country or the university\_name column feature could separate them from one another. Although I included the country column later for separate analysis to study whether it improved the accuracy or MSE of the model, but no improvement was from when country feature was not included in the dataset for analysis.
11. Scaling the features using MinMax scaler.

**Regression Tree with total\_score as Response**

When we fit regression Decision Tree model with total\_score as response, we obtain an accuracy of 0.967. After plotting the decision tree, we observe that the feature at the primary node responsible for binary split was ‘research’ with an MSE of 87.3. This implies that ‘research’ is the most important feature in the regression decision tree model which is extended to its maximum depth. At the second node, ‘citation’ and ‘research’ features are used for binary split. Test MSE of the regression tree model is 3.36.

When we perform cross validation to determine the optimal tree complexity, we check for an integer range of max\_depth = [1,14]. We notice that as we increase the depth, the accuracy increases but along with it, the cross-validation score varies. We also observe that the test MSE decreases with an increase in max depth. We obtain optimal cross validation score and accuracy at depth = 11 in the interval. Pruning the tree does not improve the test MSE but rather worsens it as we can see that the model with minimum depth has the highest test MSE when compared with models with higher tree depth.

Fitting Bagging regression model on this dataset returned an accuracy of 0.97 which is slightly better score than the regression Decision Tree model which supports our claim that ensemble models result in better models. The test MSE obtained for this model is 2.47. We observe that the ‘research’ feature has the highest importance of 0.74 followed by ‘teaching’ (0.18) and ‘citation’ (0.05).

When we fit Random Forest regression model, we observe that accuracy increases to 0.98. This is a slight improvement from the Bagging model considering the fact that the mixing of feature along with mixing of rows of features (bootstrapping) that results in a model with average of trees with lesser MSE and a better R2 value than the bagging model. The test MSE obtained for Random Forest model is 2.12. We observe that the ‘research’ feature has the highest importance in Random Forest mode too with 0.88 followed by ‘citation’ (0.05) and ‘teaching’ (0.045). When we increase the number of variables considered at each split in the integer interval [2,10), the error rate decreases for this model as when we increase the number of variables at splits, we generalize the data in a better way than those with lesser variables considered at each split.

When we fit gradient boost regressor tree, we observe that we achieve the best accuracy among all ensemble model of 0.987. This is an improvement from the Bagging and Random Forest models. The test MSE obtained for this model is also considerably lower (1.54) than the other two ensemble model. These two facts tell us that gradient boost regressor tree presents us with the best model in terms of this regression dataset.

**Classification Tree with world\_rank as Response**

When we fit classification Decision Tree model with world\_rank as response, we obtain an accuracy of 0.375. After plotting the decision tree, we observe that the feature at the primary node responsible for binary split was ‘citation’ with a Gini Index of 0.891. This implies that ‘citation’ is the most important feature in the classification decision tree model which is extended to its maximum depth. At the second node, ‘citation’ and ‘research’ features are used for binary split. Test MSE of the classification decision tree model is very high (6178500464) as this is MSE function is used for regression models and not for classification models. Maximum Likelihood Estimators are better suited for classification models.

When we perform cross validation to determine the optimal tree complexity, we check for an integer range of max\_depth = [1,6]. We notice that as we increase the depth, the accuracy increases and along with it, the cross-validation score increases. We also observe that the test MSE increases with an increase in max depth. We obtain optimal cross validation score, accuracy and MSE at depth =4 in the interval. Pruning the tree improves the test MSE but decreases the accuracy and cross validation score.

Fitting Bagging classification model with KNN classifier we obtained an accuracy of 0.386 and when we used Decision Tree classifier on this dataset we got an accuracy of 0.437 which is better score than the KNN classifier. Also, Bagging model has a better prediction accuracy than classification Decision Tree model which supports our claim that ensemble models result in better models. The test MSE obtained for this model is 5663040339 which is an improvement from Decision Tree model, but is still pretty large. We observe that the ‘citation’ feature has the highest importance of 0.32 followed by ‘research’ (0.19) and ‘teaching’ (0.11).

When we fit Random Forest classification model, we observe the accuracy to be 0.42. There is no real improvement from the Bagging model meaning the fact that the mixing of feature along with mixing of rows of features (bootstrapping) that results in a classification model with average of trees with worse MSE and R2 value than the Bagging model. The test MSE obtained for Random Forest model is 6179183964. We observe that the ‘citation’ feature has the highest importance in Random Forest mode too with 0.28 followed by ‘research’ (0.16) and ‘teaching’ (0.12). When we increase the number of variables considered at each split in the integer interval [2,10), the error rate fluctuates for this model increasing and decreasing non uniformly. We obtain the best MSE value at m = 3, i.e., if we consider to have 3 variables at each split, we get the minimum error for our Random Forest classification model.

When we fit Gradient Boost classification tree, we obtain an accuracy of 0.4. There is not much improvement from the Bagging and Random Forest models in terms of accuracy. The test MSE obtained for this model is considerably lower than the other two ensemble model. Due to the lowest MSE and comparable accuracy with other ensemble models, we can say that Gradient Boost classification model is the better model.