Problem Statement:

Given a dataset with 29 features and a class variable y which has 5 categories namely 0,1,2,3,4 for each observation in the training data. The objective is to suggest a category for the class variable using the 29 features provided for the observations.

Exploratory Data Analysis and Data Pre-processing:

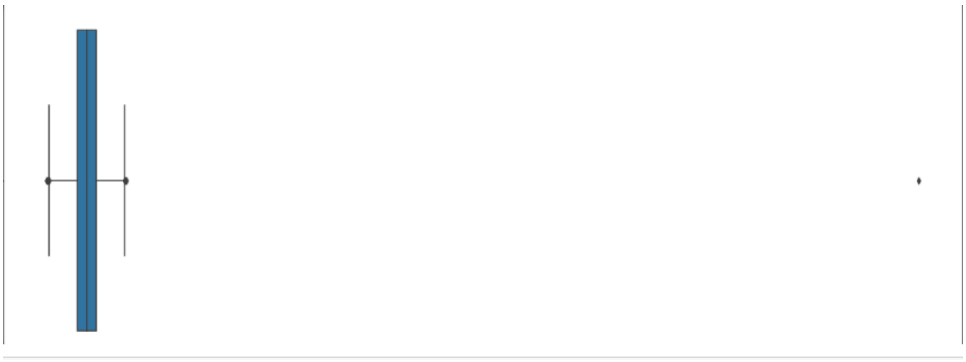
Encoding:

- Converted categorical column 26 into integers using LabelEncoder().

Outliers:

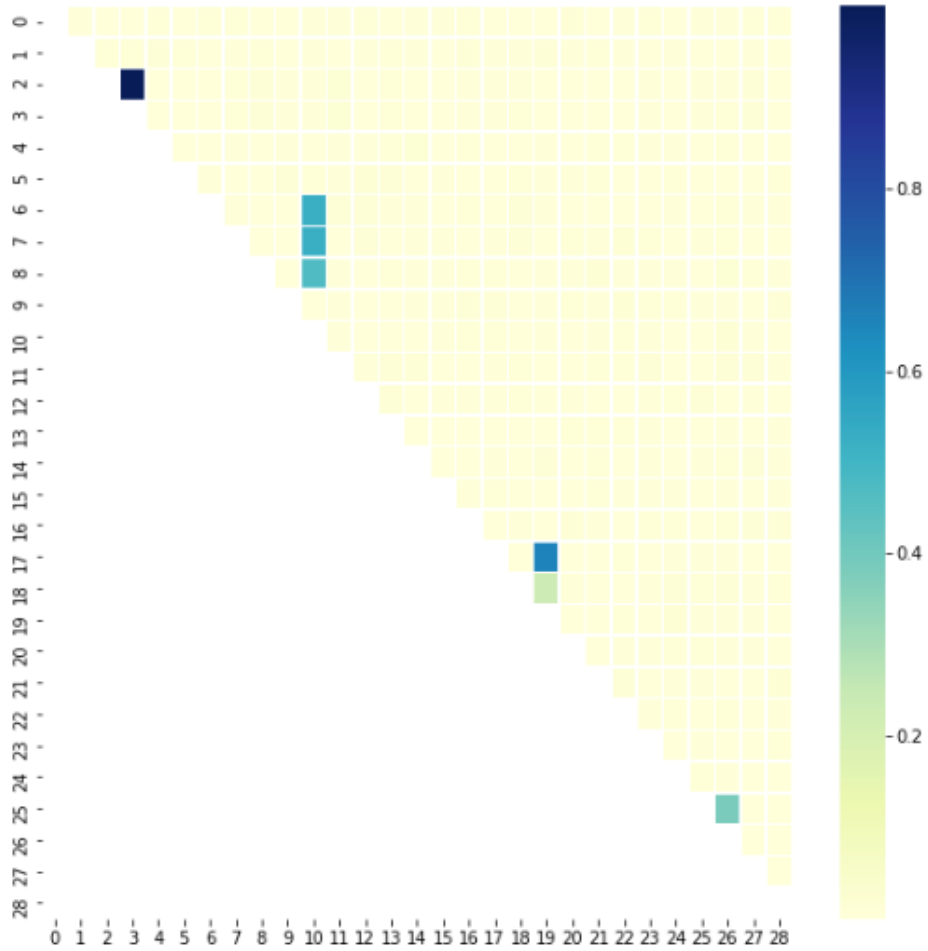
Calculated interquartile range for all the numerical columns to check for any outliers in the dataset. The values of the outliers were adjusted as follows.

- If the value is below q25-(1.5*intr_qr) where q25 is 25th percentile and intr_qr is the interquartile range, then it's adjusted to q25-(1.5*intr_qr).
- If the value is above q75-(1.5*intr_qr) where q75 is 75th percentile and intr_qr is the interquartile range, then it's adjusted to q75+(1.5*intr_qr).



Correlation:

- Removed all the rows having null values and calculated correlation between all the independent features. Used heatmap and found out that the feature 3 had collinearity greater than 80% and hence excluded from analysis.

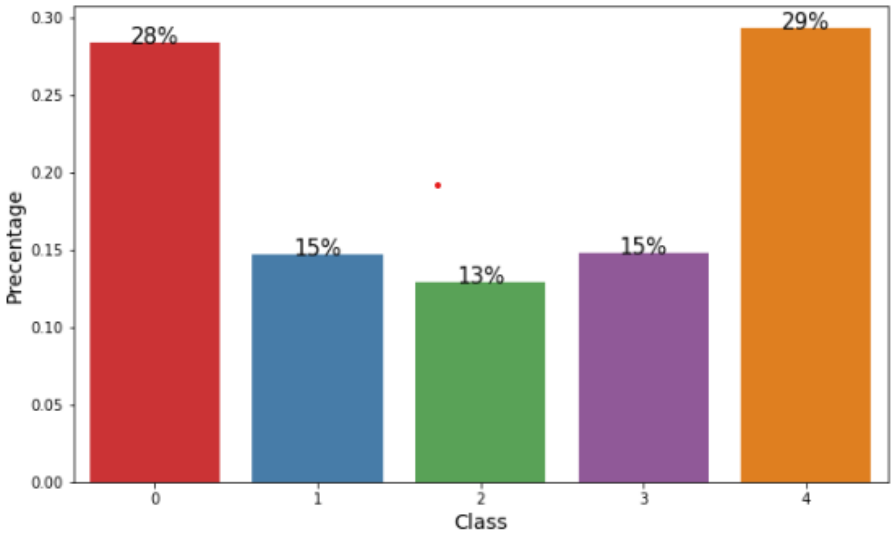


Missing value imputation:
Removed all the missing rows from analysis as there were less than 0.5%.

Normalizing the data:
Normalized the features 4,5,8,9,10,12,13,15,18,19,20,21,23,26,27 which had higher values to bring the data to a uniform scale.

Metric Details:

I have chosen F1 weighted average score as there was an imbalance in the class variables.



Model Details:

I have chosen the following models as it is a regression problem.

- **Logistic regression:**
  - Applied GridSearch with 5-fold cross validation to find out the best hyper parameters such as penalty, max_iter and C.
  - Applied RFE algorithm to check feature significance and removed insignificant features from the final model evaluation.
  - Calculated F1 weighted average for the dataset.

Before HyperParameter Tuning	0.4521
After HyperParameter Tuning	0.706

- **Random Forest Classifier:**
  - Applied GridSearch to find out the best hyper parameters such as criterion, max_depth, min_samples_leaf, ccp_alpha, warm_start and class_weight.
  - Applied RFE algorithm to check feature significance and removed insignificant features from the final model evaluation.
  - Calculated F1 weighted average for the dataset.

Before HyperParameter Tuning	0.422
After HyperParameter Tuning	0.706

- **Support Vector Classifier:**
  - Applied GridSearch to find out the best hyper parameters such as kernel, degree, decision_function_shape,coef0, C and class_weight.
  - Calculated F1 weighted average for the dataset.

Before HyperParameter Tuning	0.624
After HyperParameter Tuning	0.680

Conclusion:

Logistic regression has the highest weighted F1 score for this data.