**Data Analysis:**

Initially, the dataset ‘nba2021.csv’ consisted of 28 attributes (including the target attribute). To decrease the dimensionality and computations time of the problem, data analysis was carried out selectively drop attributes. Attributes such as “Player”(Player name), “Age” and “Tm”(Team) either has too many attribute values or were not as relevant to the problem definition; therefore, they were dropped.

Next, the class distribution of the target attribute was calculated:

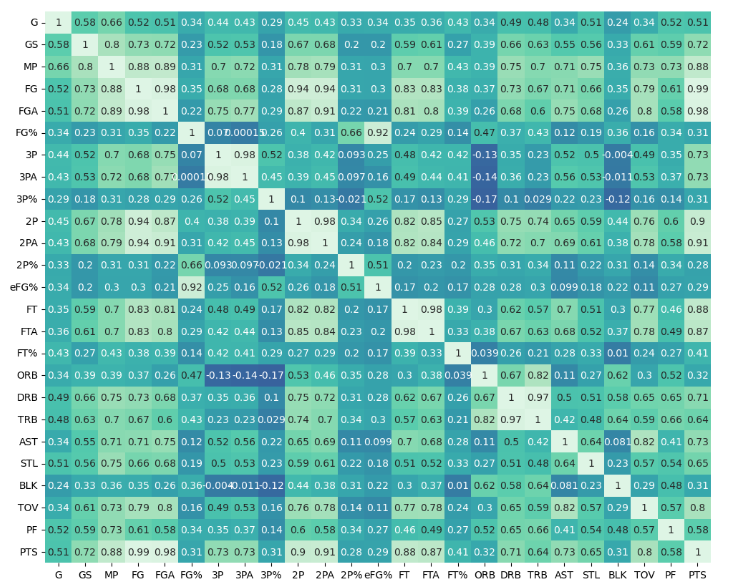
Table

Description automatically generated

Chart

Description automatically generated

To find if a linear relationship exists between various attributes, correlation matrix was obtained and viewed as a heatmap:

­

The above heatmap shows that features such as: FG, FGA; 3P, 3PA; 2P, 2PA and FT, FTA were found to be highly correlated. For this assignment, all the "attempts" columns were kept and attributes like ‘2P’, ‘2P%’ were dropped. In addition, ’DRB’ had a high correlation with 'TRB'. Therefore, one of them will be dropped.

Analysis on points:

Text

Description automatically generatedText

Description automatically generatedChart, line chart

Description automatically generatedUpon analyzing the distribution of point-related attributes such as: 3PA, 2PA, FGA, FTA, eFG% and PTS with respect to the target attribute (Position), the distribution for ‘FTA’ (Free Throws Attempted) and ‘2P’ was found to be similar for all positions. Therefore, they were dropped:

Table

Description automatically generatedChart, line chart

Description automatically generated

Analysis on G, GS, MP: The distributions of Games Played and Games Started were also found to be similar for all positions and dropped:

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

All remaining attributed were also analyzed in a similar fashion and ‘TOV’, ‘PF’ were also dropped. In addition, the difference between the average number of minutes played by players in each position were not found to be too different; therefore ‘MP’ was also dropped.

Chart, line chart

Description automatically generatedTable

Description automatically generated with low confidence

Table

Description automatically generatedChart, line chart

Description automatically generated

Table

Description automatically generatedChart, line chart

Description automatically generated

The remaining attributes after were:

A screenshot of a computer

Description automatically generated with low confidence

**2) Model Selection**

‘train\_test\_split’ module from sklearn was used to split the data with stratification. 5 different types of classification models were trained and tested to obtain the following results:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Test Accuracy** | | | | | | | | | |  |
| **Model** | **Trial 1** | **Trial 2** | **Trial 3** | **Trial 4** | **Trial 5** | **Trial 6** | **Trial 7** | **Trial 8** | **Trial 9** | **Trial 10** | **Average** |
| **Decision Tree** | 0.48 | 0.4 | 0.43 | 0.44 | 0.456 | 0.52 | 0.456 | 0.416 | 0.425 | 0.46 | 0.4483 |
| **Random Forest** | 0.576 | 0.496 | 0.52 | 0.552 | 0.52 | 0.432 | 0.51 | 0.53 | 0.5 | 0.505 | 0.5141 |
| **Logistic Regression** | 0.56 | 0.504 | 0.56 | 0.544 | 0.568 | 0.56 | 0.52 | 0.54 | 0.53 | 0.52 | 0.5406 |
| **SVM** | 0.576 | 0.52 | 0.536 | 0.536 | 0.552 | 0.56 | 0.48 | 0.52 | 0.51 | 0.5 | 0.529 |
| **KNN** | 0.52 | 0.496 | 0.528 | 0.48 | 0.512 | 0.496 | 0.48 | 0.49 | 0.5 | 0.47 | 0.4972 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.N. | Model | Best  Hyperparameters  (test) | Best Training Accuracy | Best test Accuracy | Overfitting  observed |
| 1 | Decision Tree | Depth = 5 | 1 | ~ 0.48 | Yes |
| 2 | Random Forests | Depth = 12 | 1 | ~0.57 | Yes |
| 3 | Logistic Regression | C = 2 | ~0.56 | ~0.57 | No |
| 4 | SVM | C = 1 | ~0.62 | ~0.576 | No |
| 5 | KNN | K = 8 | 1 | ~0.528 | Yes |

The best average test accuracy was obtained for Logistic Regression with a regularization parameter of C = 2; therefore, it was used as the final model for cross validation process. Low training accuracy for logistic regression and SVM suggests that all classes in the data may not have been linearly separable. A high training accuracy of 1 for tree-based methods and KNN suggests overfitting on the training data.

**3) Confusion Matrix (Logistic Regression)**

Upon a number of training iterations, the logistic regression model was found to have an overall higher accuracy on average and chosen as the final model for prediction and cross validation:

Calendar

Description automatically generated

A picture containing calendar

Description automatically generated

**4) 10 – Fold Cross Validation**

The entire dataset was used for cross validation. Since logistic regression was being used as the classifier, the data was scaled using MinMaxScaler() from sklearn.preprocessing.

Model Hyperparameters: C = 2; max\_iter = 500

**5) Accuracy Scores**

**Table

Description automatically generated**

The average score was 0.537.