U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Data Mining Project

Aprendizagem Computacional (AC) - 2023

Grupo G64

Anete Pereira - up202008856@edu.fe.up.pt - 33,3%
Hugo Castro - up202006770@edu.fe.up.pt - 33,3%
José Araújo - up202007921@edu.fe.up.pt - 33,3%

Domain description

WNBA Competition Structure

- Regular season followed by playoffs.
- Teams aim for the playoffs by winning games.

Dataset Overview

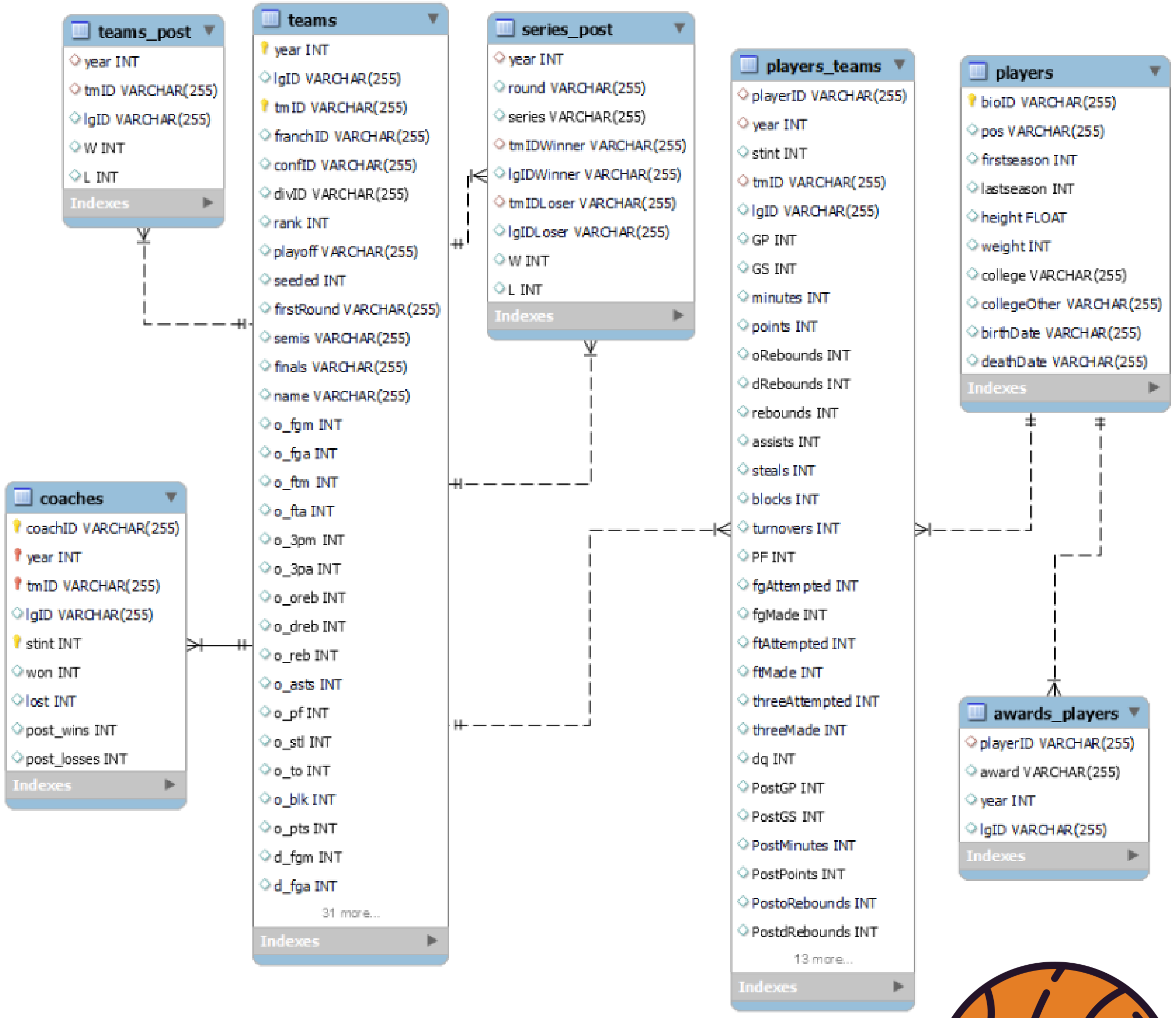
- 10 years of data on players, teams, coaches, and game metrics.

Dataset Description

- 57 coaches, 893 players, 10 seasons, 20 teams

Project Goal

- Use machine learning to predict which teams will qualify for the playoffs.



Exploratory Data Analysis (1/3)

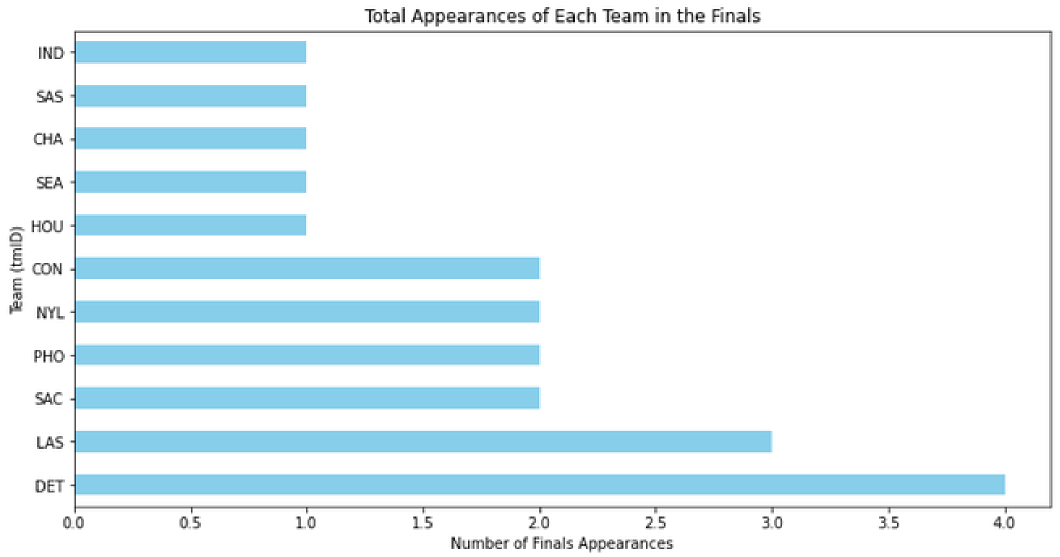


Fig.3 - teams that went to final rounds in the playoffs and won or lost.

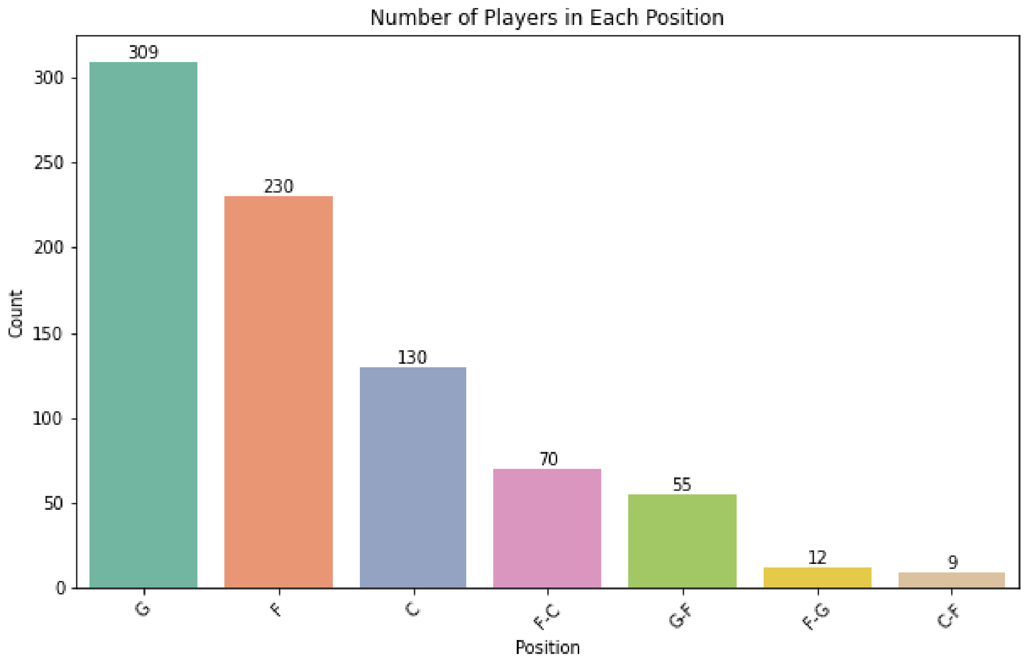


Fig.4 - number of players in each position

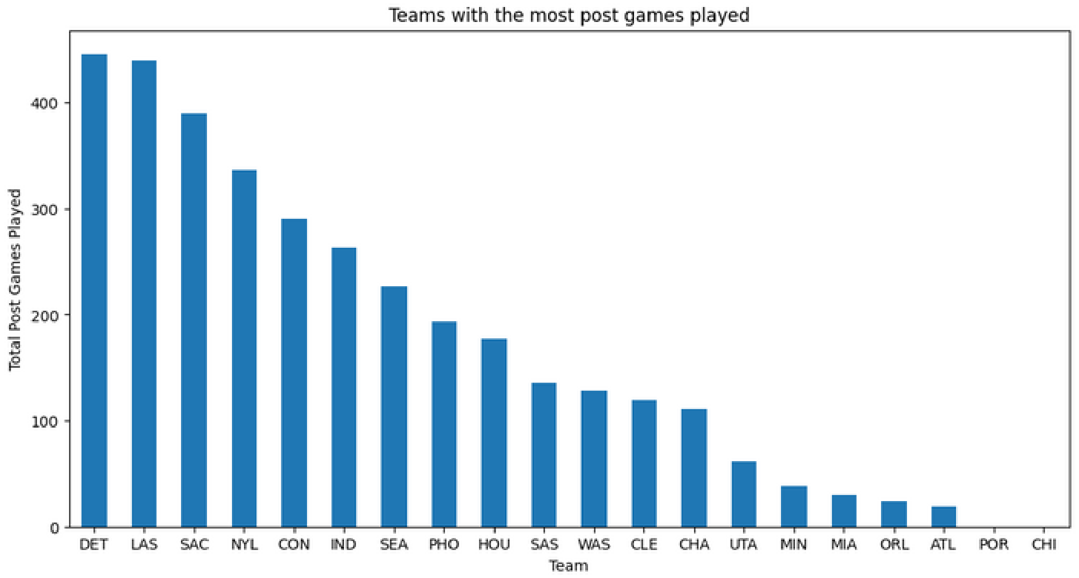


Fig.5 - teams with the most games played

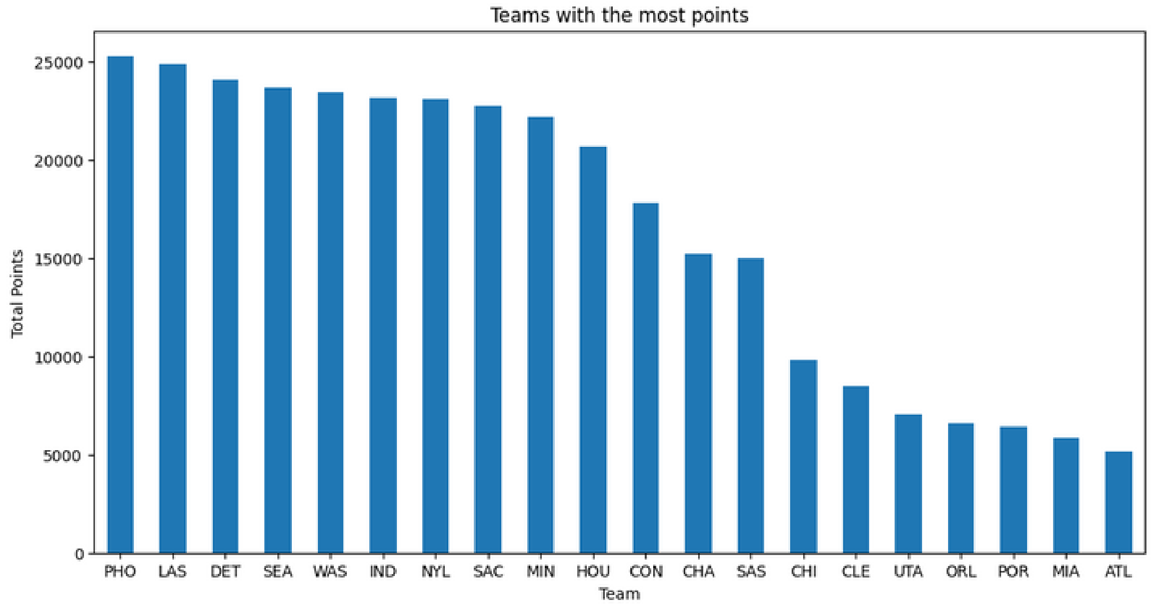


Fig.6 - teams with the most points

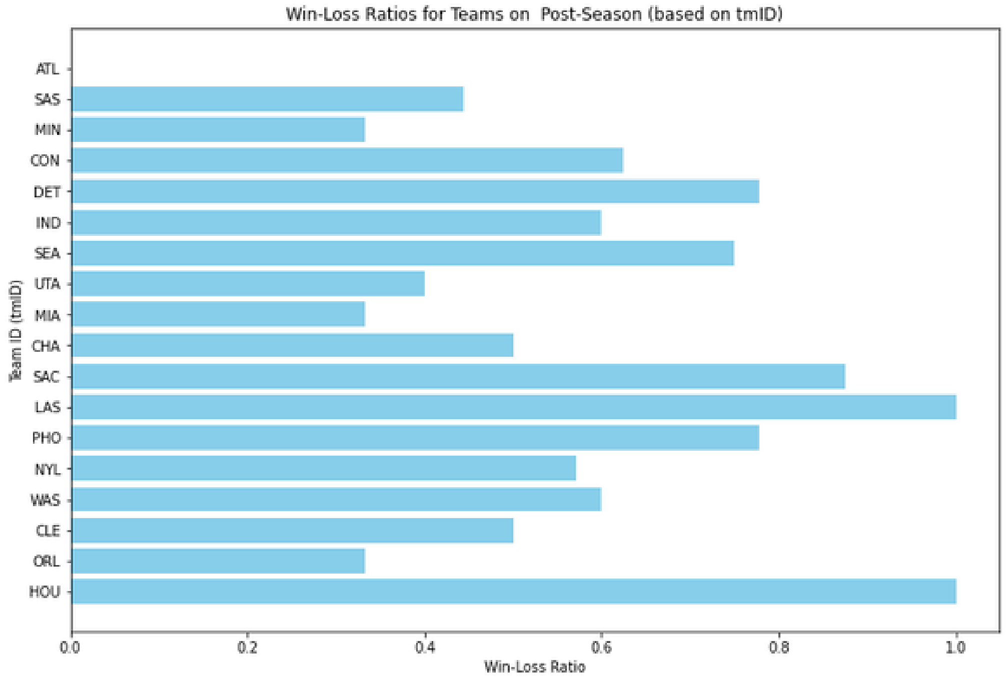


Fig.7 - win-loss ratio of teams during post season

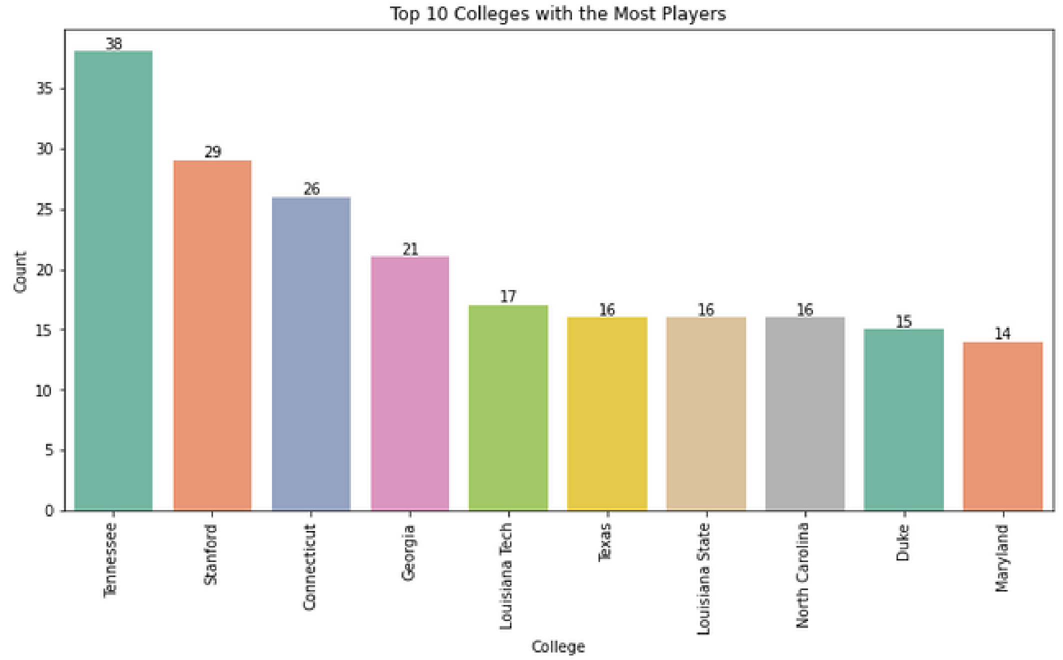


Fig.8 - top 10 colleges from where players come from

Exploratory Data Analysis (2/3)

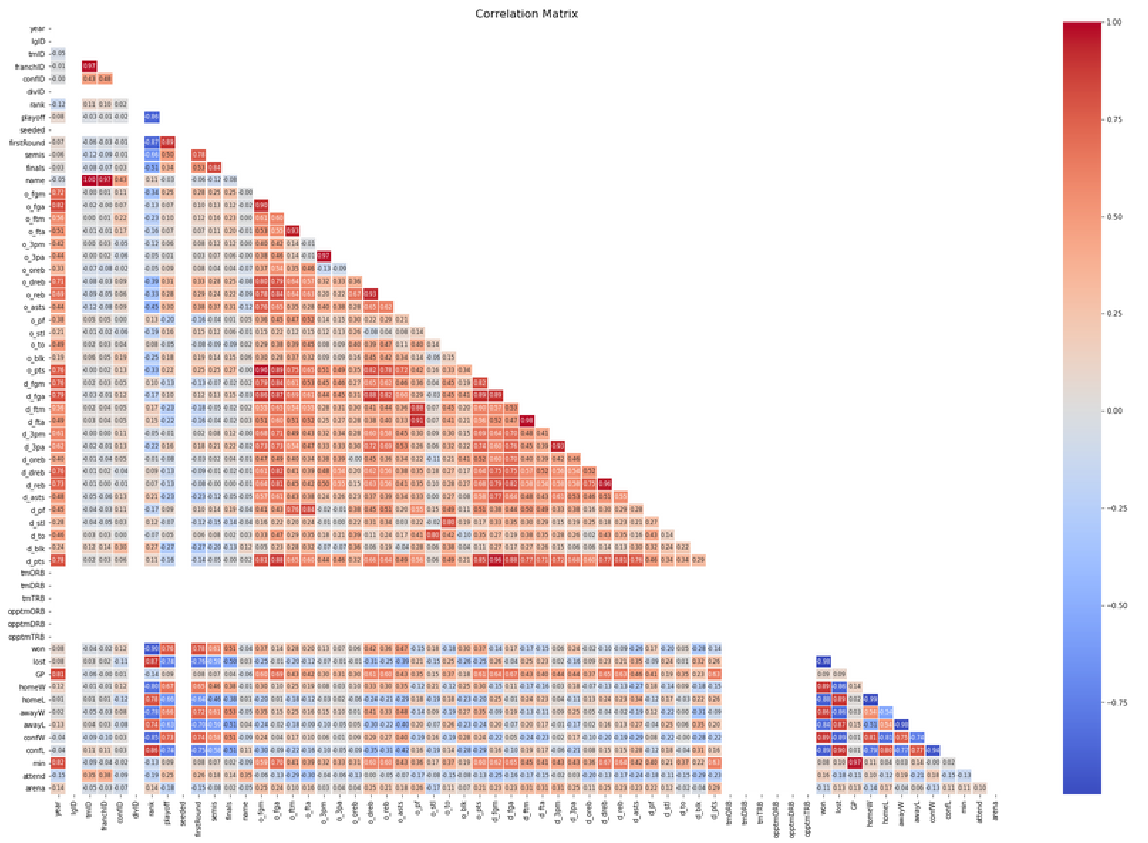


Fig.9 - teams table correlation matrix

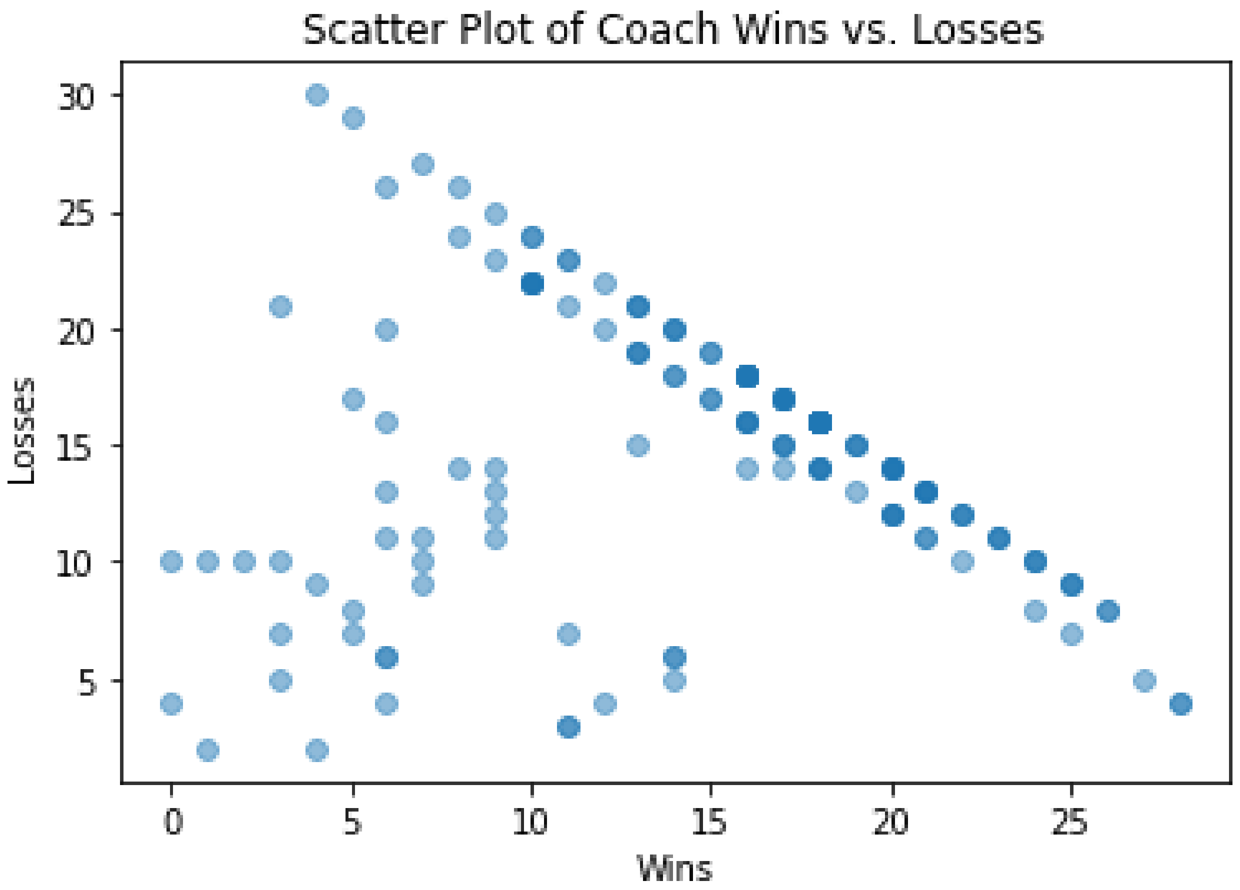


Fig.10 - relation between coaches wins and losses

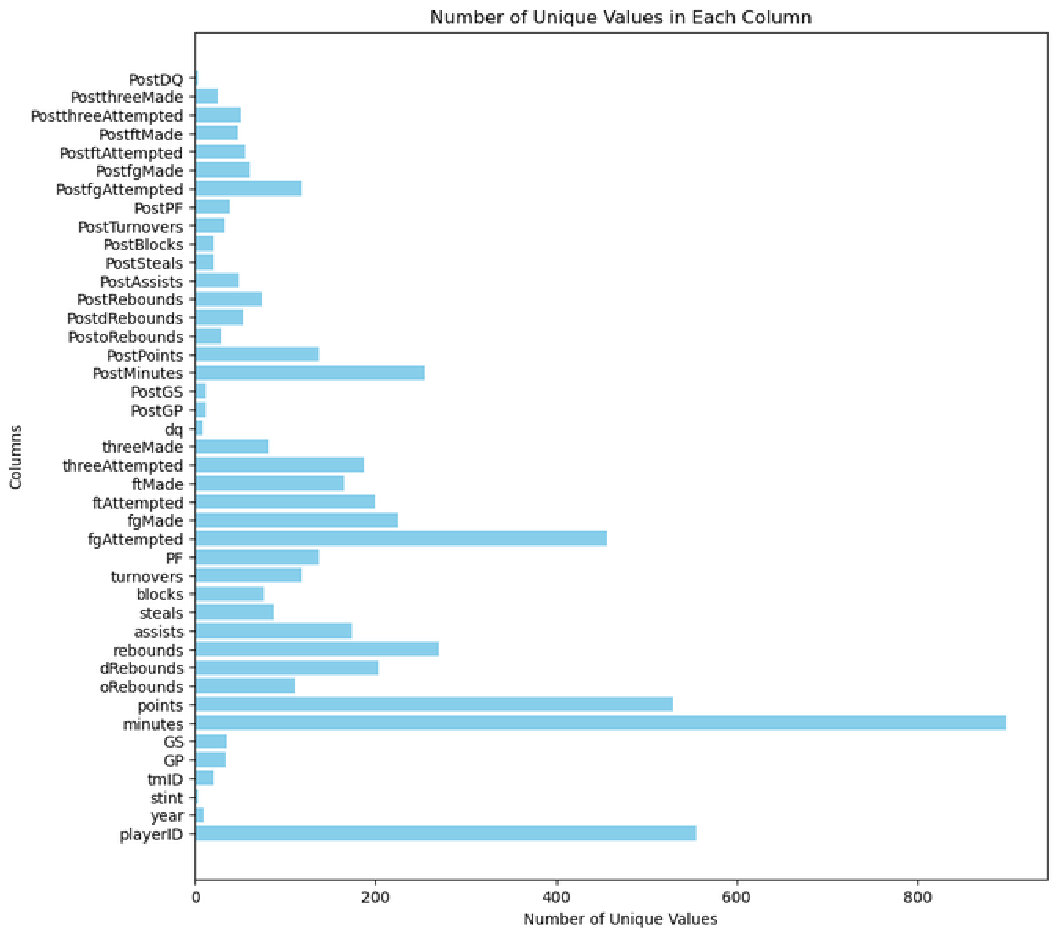
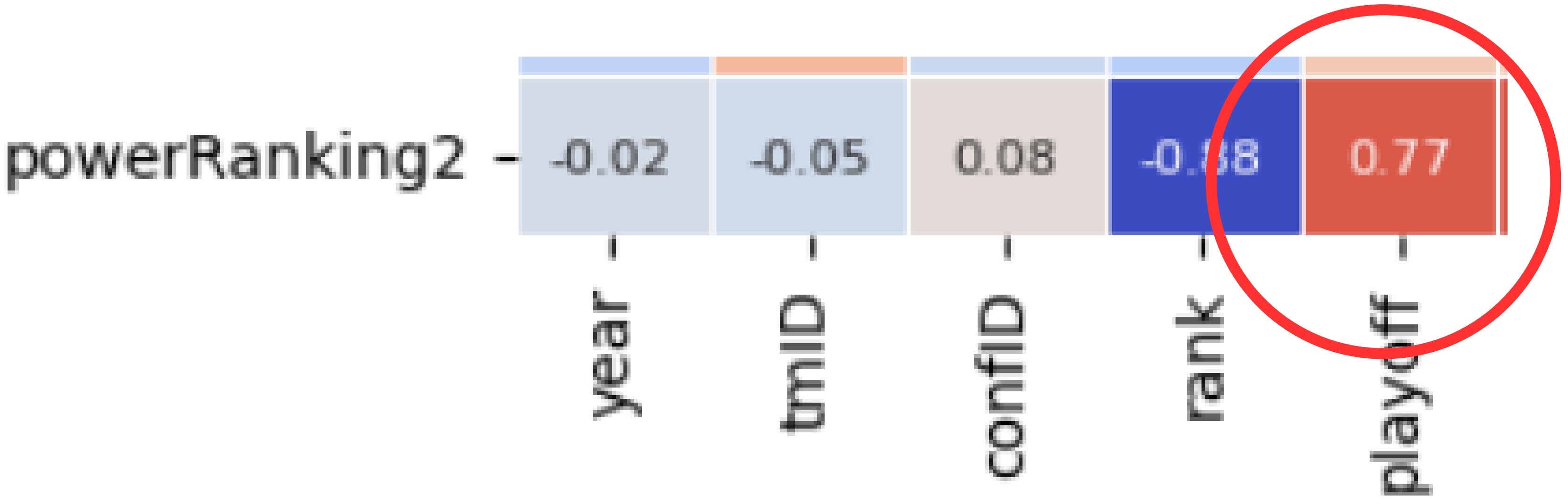
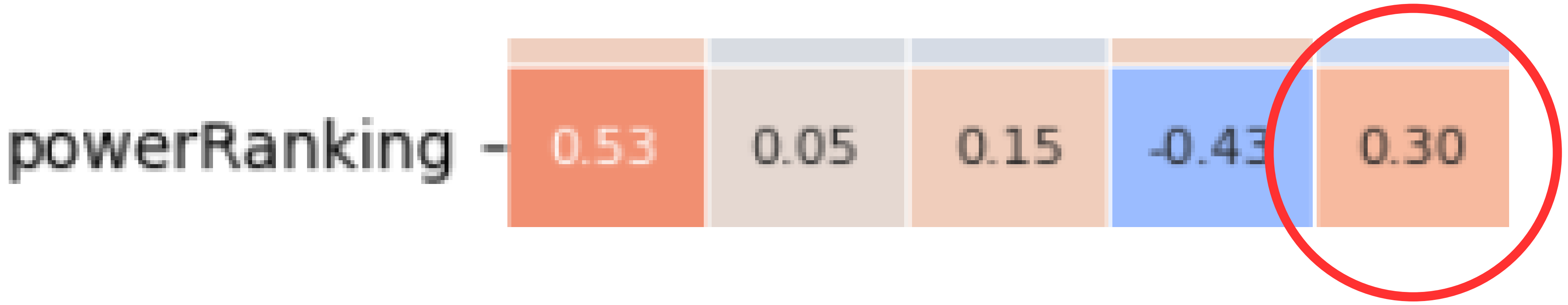


Fig.11 - number of unique value in each column of teams table

Exploratory Data Analysis (3/3)






```
chi_square(team2, 'playoff')
[31] ✓ 0.1s

Chi-square test for year and playoff:
Chi-square value: 1.3666430343849703
P-value: 0.9980204687073957
Fail to reject the null hypothesis. There is not enough evidenc

Chi-square test for tmID and playoff:
Chi-square value: 32.49763248847927
P-value: 0.027449727442441722
Reject the null hypothesis. There is a significant association.

Chi-square test for confID and playoff:
Chi-square value: 0.00046010944700461995
P-value: 0.9828865593186417
Fail to reject the null hypothesis. There is not enough evidenc

Chi-square test for rank and playoff:
```

```
point_biserial(team2, team2.columns, 'playoff')
[32] ✓ 0.0s

                      correlation       p_value
year                     0.077051  3.620822e-01
tmID                    -0.025299  7.650471e-01
confID                  -0.016001  8.500968e-01
rank                          NaN           NaN
playoff                  1.000000  0.000000e+00
firstRound               0.892829  2.398164e-50
semis                    0.497063  3.130200e-10
finals                   0.336818  4.156818e-05
o_oreb                  -0.092363  2.742871e-01
o_dreb                  -0.307557  1.966872e-04
d_oreb                   0.075642  3.709516e-01
d_dreb                   0.128503  1.274875e-01
min                      0.087890  2.982981e-01
arena                   -0.178068  3.399359e-02
powerRanking2            0.778807  3.781603e-30
```

Problem Definition:

In each season the competition consists in two distinct phases.

During the initial phase, all teams compete against one another with the goal of achieve the greatest number of wins possible.

After that phase, a predetermined selection of teams that have achieved the most wins qualifies for the playoff stage.

Objective

Predict which teams will qualify for the playoffs in the next season.

Data

Players, teams, coaches, games and other metrics data from 10 years.

Success Criteria

Evaluate model performance resorting to metrics like accuracy, recall, f1-score, precision, etc.

Data preparation (1/4)

Feature Selection

Remove **redundant** attributes: erased **lgID** from all tables because it's the same (WNBA)

- teams: **erase** 'lgID', 'divID', 'seeded', 'tmORB', 'tmDRB', 'tmTRB', 'opptmORB', 'opptmDRB', 'opptmTRB'...
- players: **erase** 'collegeOther', 'deathDate', 'firstseason', 'lastseason' - lot of missing values.
- coaches: **erase** 'lgID', 'post_wins', 'post_losses'

Join tables based on ID's

Join tables based on ID's to create

- **join** players, player_teams, teams, coaches and award_players on **playerID, tmID and year**
- pay attention to type of joins (inner, outer) so that the data is not badly joined
- sometimes we may need to merge datasets using multiple keys

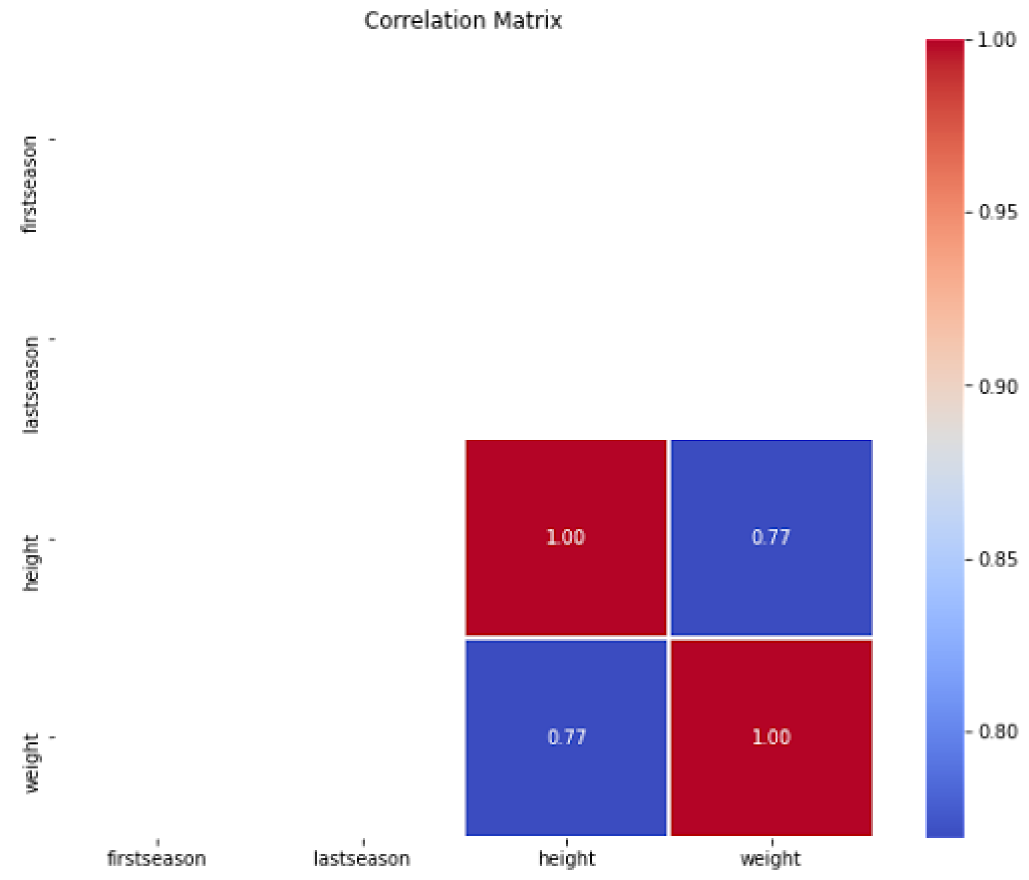


Fig.12 - correlation matrix between data erased in players table

Data preparation (2/4)

Feature Engineering

Generate new features to improve the performance of the machine learning model.

- generate column with **Power Ranking** for teams, players and coaches
- extract age from player birthdate
- divide table awards_players into awards_players and awards_coaches
- generate column with count of player awards

Outlier Detection

Algorithms used to detect outliers

- Z-Score (we did not notice any relevant outlier, all values make sense given our problem)

Data preparation (3/4)

Add coaches and players awards as team metrics

For coaches:

- Verify if the coach has a “coach of the year” award and add to column coachOfYear of coaches dataset.
- When calculating the coaches power ranking, add 50 extra points if the coach was considered coach of the year

For players:

- Verify if the player has one of these awards: “Defensive Player of the Year, Most Improved Player, Most Valuable Player, Rookie of the Year, Sixth Woman of the Year, WNBA Finals Most Valuable Player”
- Count the number of awards the players of a team has in a year and store it in column playersAwards of feature engineering dataset

Data preparation (4/4)

Inconsistency in the dataset:

Comparing assists and points of teams with sum of assists and points of players of that year

For example:

tmID: **HOU**
Ano: 7

team offensive points: 2507
team offensive assists: 532

sum players points: 2428
sum players assists: 516

Experimental setup (1/5) - Pipeline

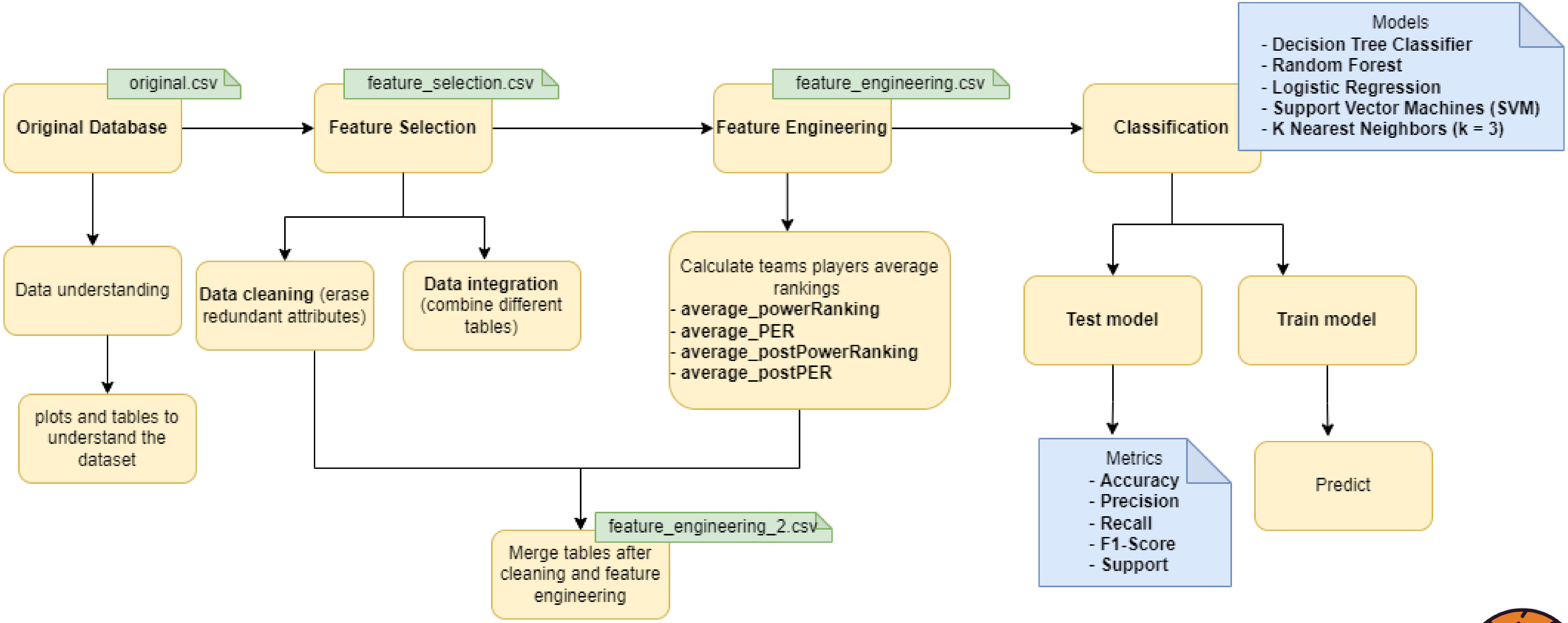


Fig.12 - Experimental setup pipeline

Experimental setup (2/5) - Player and Team Ranking

- **Collective PowerRank**

  Performance metric calculated based on metrics available in teams dataset

- **Average of individual PowerRanks**

  Performance metric calculated based on the average of individual player powerranks.

- **PER**

  Performance metric used in real life to measure a player's overall performance by considering various stats and putting them in context.

Experimental setup (3/5) - Power Ranks

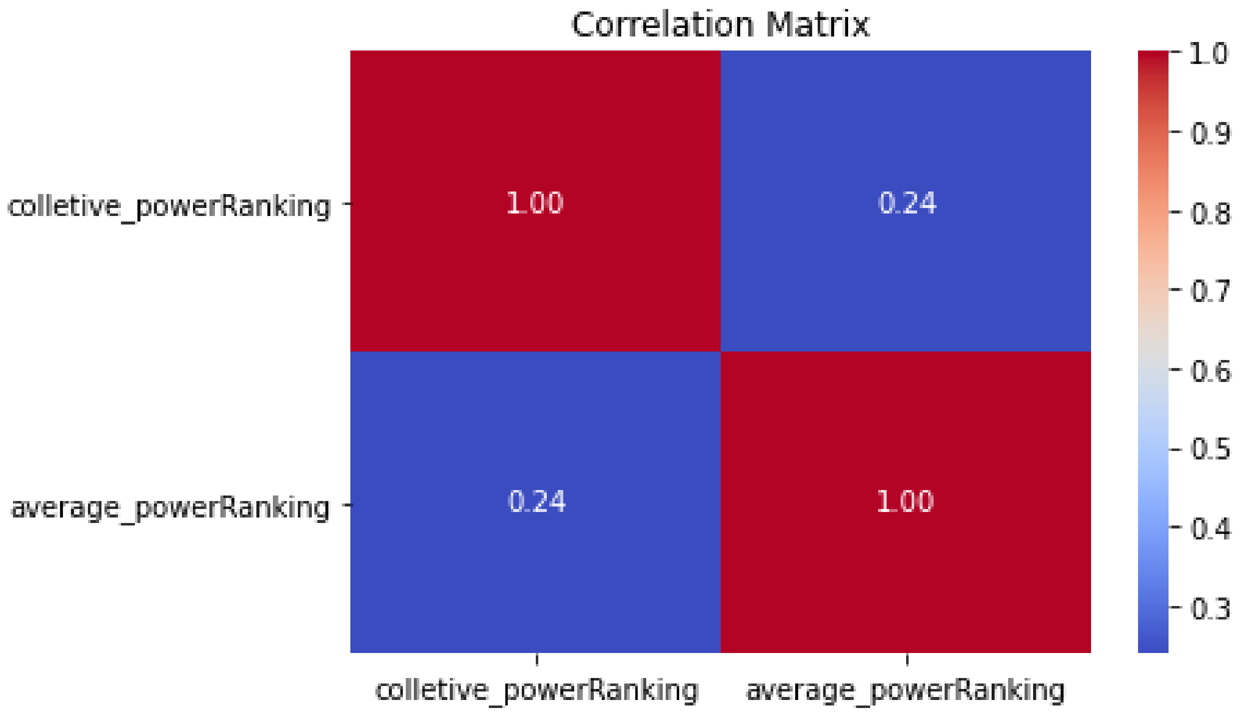


Fig.13 - Correlation Matrix between the team collective and average of individual power rankings

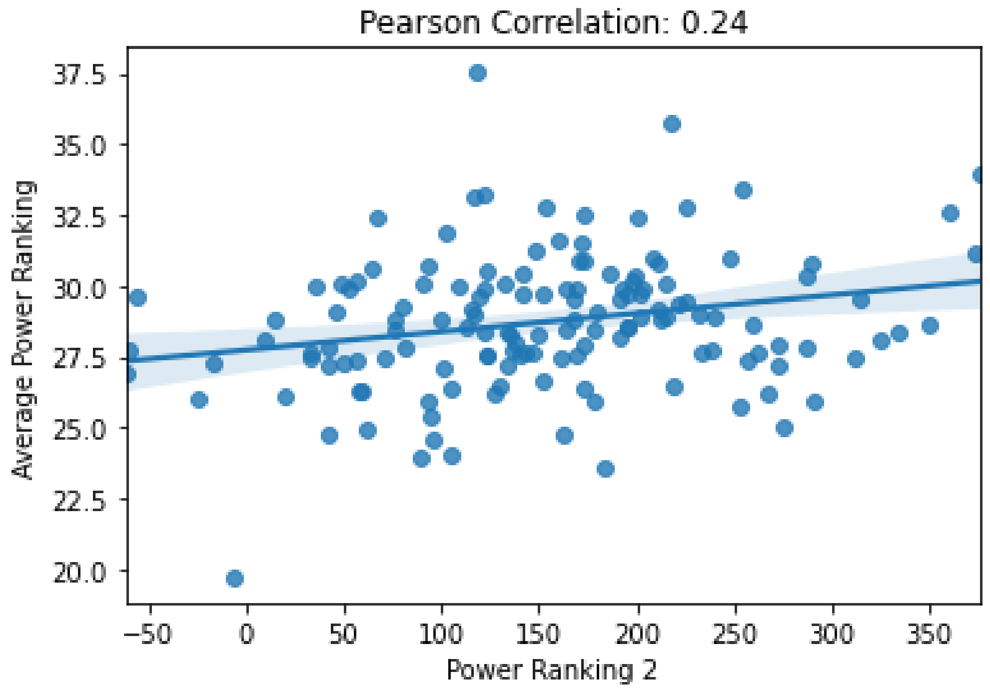


Fig.14 - Pearson correlation between the team collective and average of individual power rankings

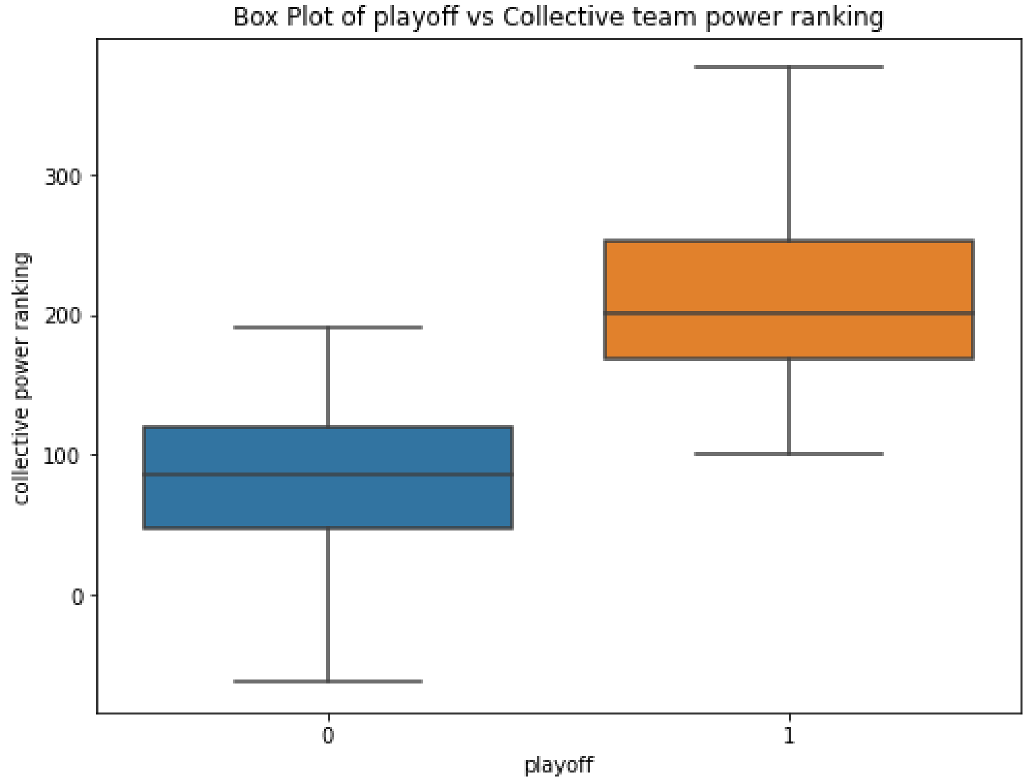


Fig.15 - Box plot comparing playoff with the collective power rank

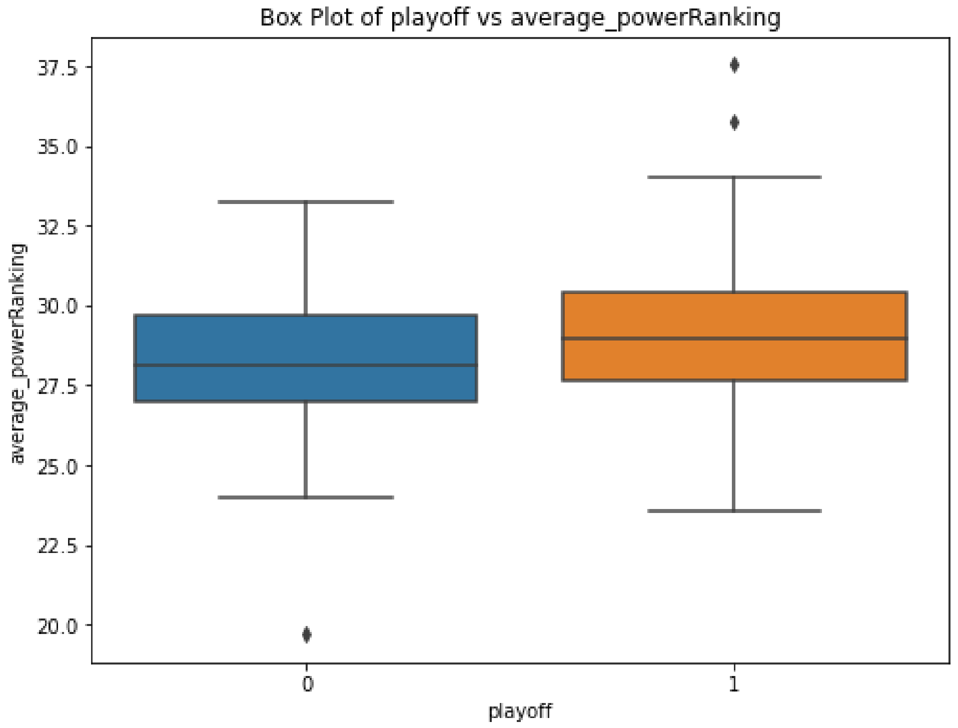


Fig.16- Box plot comparing playoff with the average of individual power ranks

Experimental setup (4/5) - Sliding Window

- Used to **understand how teams perform over different years**, can capture temporal patterns and account for changes
- Helps the model learn from the past and **adapt to how teams play as time goes on.**
- Keeps predictions relevant by **considering recent team trends instead of relying on outdated information**.
- Ensures that the predictions work well across different years, making them more reliable.

Experimental setup (5/5) - Other metrics

Hyperparameter Tuning

- Uses GridSearchCV to perform a grid search cross-validation to find the best hyperparameters based on accuracy; Creates a classifier with the best parameters and fits it to the training data

Custom Binary Classification

- Creates a binary classification for playoff selection based on a threshold
- Plots a learning curve for the models on the training set

ROC Curve and AUC

- Plots the ROC curve and calculates the Area Under the Curve (AUC) for the model on the test set

Results (1/4) - Chosen models

K-Nearest Neighbors (K=3)

- Classifies or predicts a data point based on the majority class or average value of its k nearest neighbors
- Shows no clear signs of overfitting
- Demonstrates learning and adaptability over different years
- Accuracy ranges from 54% to 62%, indicating consistent performance

Logistic regression

- Utilizes the logistic function to model the relationship between independent variables and the outcome probability
- Provides stability and reliability in playoff prediction
- Demonstrates balanced precision, recall, and F1-score metrics
- Achieves accuracy levels between 62% and 85% over different years

Results (2/4) - KNN

TeamID	Probability	confID
IND	1.000000	0
WAS	0.803922	0
NYL	0.796187	0
ATL	0.786089	0
LAS	1.000000	1
SEA	0.806747	1
SAS	0.805443	1
PHO	0.755461	1

Results (3/4) - Logistic Regression

TeamID	Probability	confID
IND	0.686146	0
DET	0.642155	0
ATL	0.609116	0
WAS	0.544017	0
PHO	0.763195	1
LAS	0.708733	1
SEA	0.532808	1
SAS	0.483729	1

Results (4/4) - Final Prediction for Year 11

TeamID	confID
IND	0
NYL	0
ATL	0
WAS	0
PHO	1
LAS	1
SEA	1
SAS	1

Conclusions, limitations and future work

- Good Exploratory Data Analysis
- Good Feature Engineering
- Good performance in two models
- Model tuning could be better in some models
- After model performance analysis, we concluded that if the dataset was bigger we could understand model behavior better

Annexes

Annex

Coaches and Teams Outliers Detection

```
Coaches Dataset Outliers
Outliers Won: [0, 28, 28, 1, 1, 0]
Outliers Lost: [2, 29, 26, 27, 2, 30, 26, 3, 3]
Outliers Post Wins: [6, 6, 6, 6, 6, 7, 6, 7, 6, 7, 7, 7, 6]
Outliers Post Losses: [4, 4, 4, 4, 4, 5, 5]
```

```
Teams Dataset Outliers
Outliers o_fgm: [1089, 647, 671, 1079, 1063, 1069, 1128, 667]
Outliers o_fga: [2428, 2434, 2419, 2485, 2454]
Outliers o_ftm: [652, 336, 642, 643, 643, 333, 668]
Outliers o_fta: [882, 469, 864, 844, 478, 839, 827]
Outliers o_3pm: [62, 62, 259, 254, 265, 283, 257, 256]
Outliers o_3pa: [209, 205, 695, 706, 710, 722, 802, 701, 739]
Outliers o_oreb: [242, 418, 452, 246]
Outliers o_dreb: [926, 931, 906, 537]
Outliers o_reb: [1286, 1311, 1282, 793]
Outliers o_asts: [630, 640, 683, 390]
Outliers o_pf: [796, 532, 530, 509, 467, 490, 794, 784]
Outliers o_stl: [336, 354, 373, 193, 187]
Outliers o_to: [408, 633, 613, 612, 637]
Outliers o_blk: [216, 63, 179, 181, 178]
Outliers o_pts: [2861, 1831, 2960, 3025, 3010, 3156, 1822]
Outliers d_fgm: [664, 679, 1041, 691, 1036, 1094]
Outliers d_fga: [2526, 2460, 2582]
Outliers d_ftm: [679, 328, 632, 331, 325, 630, 694, 638, 347, 635]
Outliers d_fta: [918, 444, 452, 448, 852, 932, 851, 467, 851]
```

Annex - Logistic Regression - Results

Accuracy: 0.54

	precision	recall	f1-score	support
N	0.33	0.20	0.25	5
Y	0.60	0.75	0.67	8
accuracy			0.54	13
macro avg	0.47	0.47	0.46	13
weighted avg	0.50	0.54	0.51	13

Original Dataset

Accuracy: 0.54

	precision	recall	f1-score	support
N	0.33	0.20	0.25	5
Y	0.60	0.75	0.67	8
accuracy			0.54	13
macro avg	0.47	0.47	0.46	13
weighted avg	0.50	0.54	0.51	13

Feature Selection Dataset

Accuracy: 0.54

	precision	recall	f1-score	support
N	0.40	0.40	0.40	5
Y	0.62	0.62	0.62	8
accuracy			0.54	13
macro avg	0.51	0.51	0.51	13
weighted avg	0.54	0.54	0.54	13

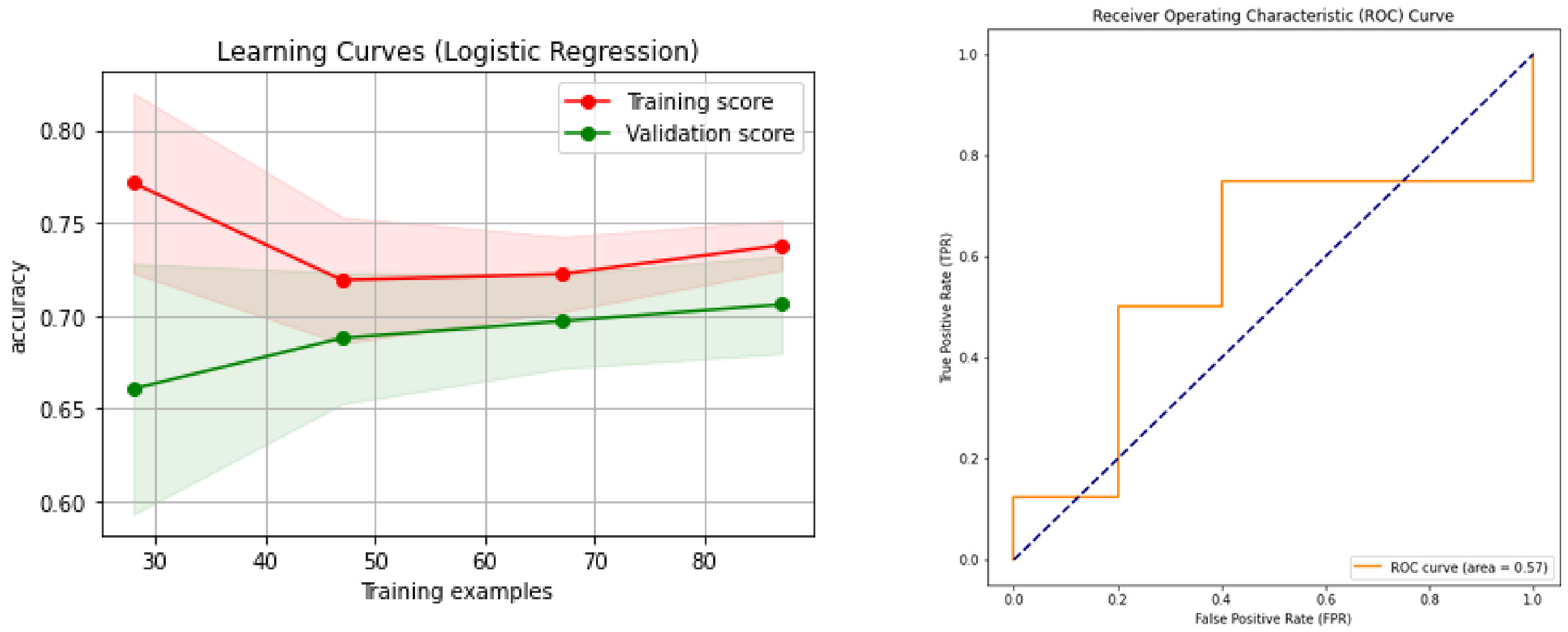
Feature Engineering Dataset

Accuracy: 0.62

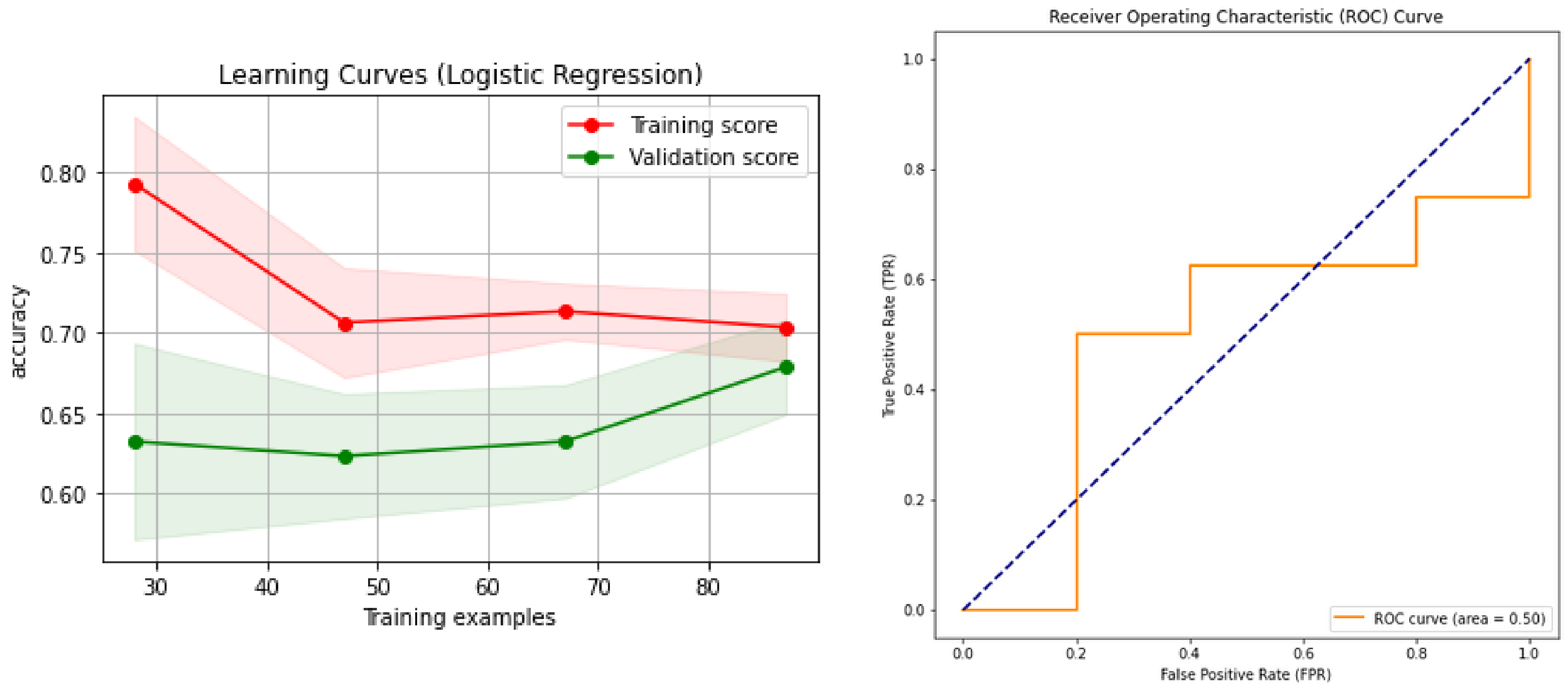
	precision	recall	f1-score	support
N	0.50	0.60	0.55	5
Y	0.71	0.62	0.67	8
accuracy			0.62	13
macro avg	0.61	0.61	0.61	13
weighted avg	0.63	0.62	0.62	13

Feature Engineering 2 Dataset
(year 9 prediction)

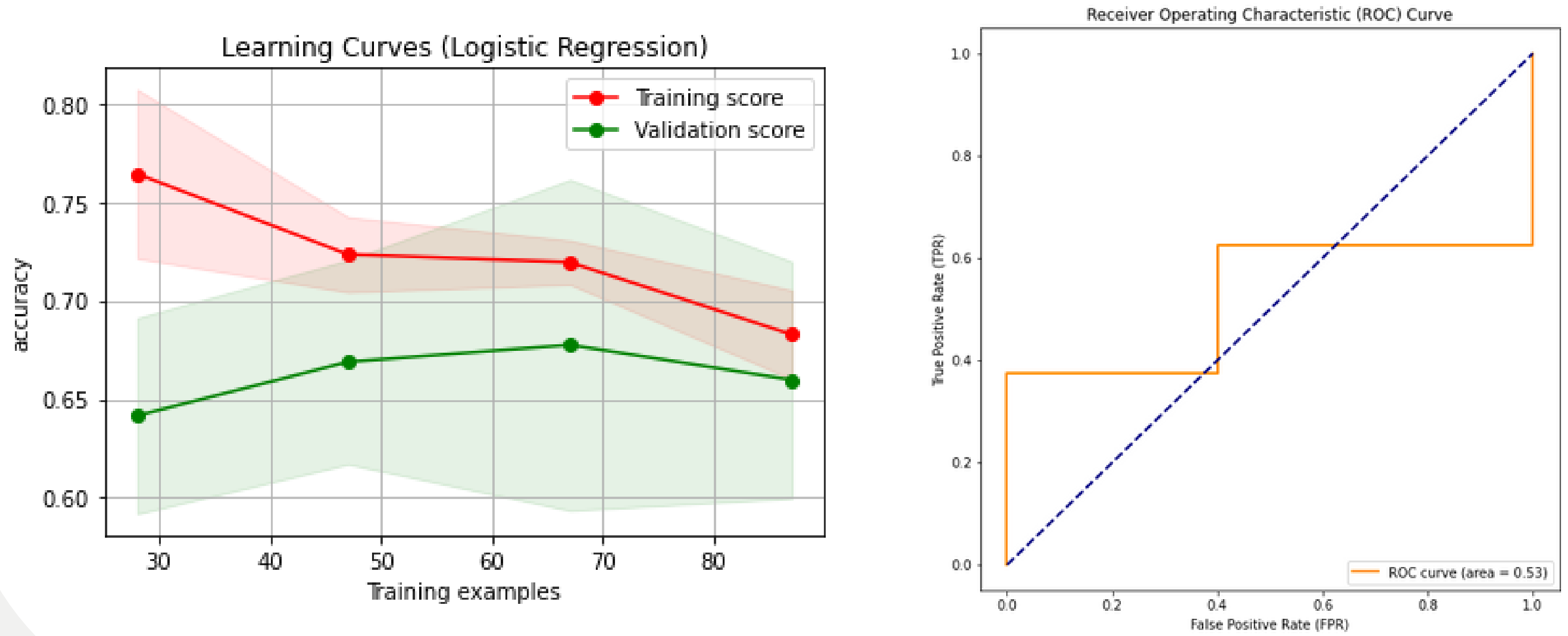
Annex - Logistic Regression- Curves



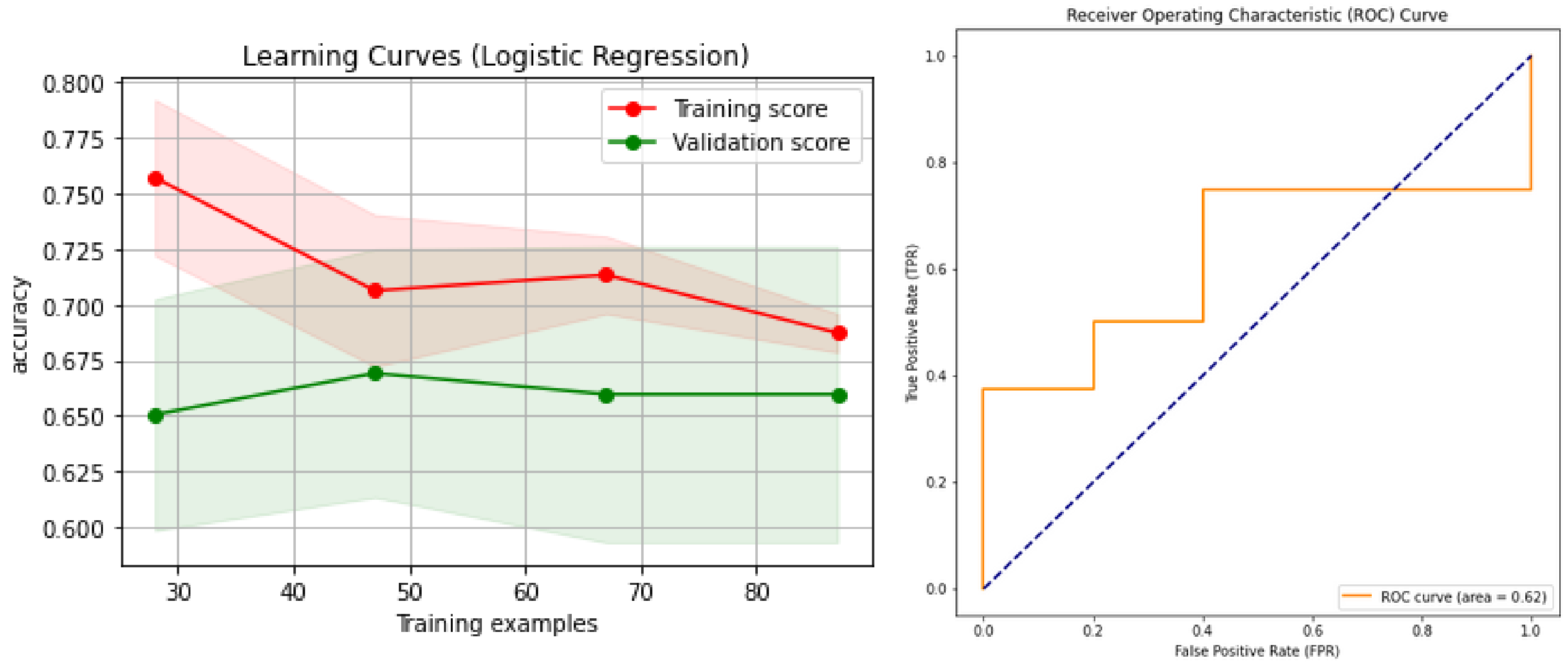
Original Dataset



Feature Engineering Dataset



Feature Selection Dataset



Feature Engineering 2 Dataset (year 9)

Annex - K Nearest Neighbors - Results

Accuracy: 0.54

	precision	recall	f1-score	support
N	0.40	0.40	0.40	5
Y	0.62	0.62	0.62	8
accuracy			0.54	13
macro avg	0.51	0.51	0.51	13
weighted avg	0.54	0.54	0.54	13

Original Dataset

Accuracy: 0.46

	precision	recall	f1-score	support
N	0.25	0.20	0.22	5
Y	0.56	0.62	0.59	8
accuracy			0.46	13
macro avg	0.40	0.41	0.41	13
weighted avg	0.44	0.46	0.45	13

Feature Selection Dataset

Accuracy: 0.62

	precision	recall	f1-score	support
N	0.50	0.40	0.44	5
Y	0.67	0.75	0.71	8
accuracy			0.62	13
macro avg	0.58	0.57	0.58	13
weighted avg	0.60	0.62	0.61	13

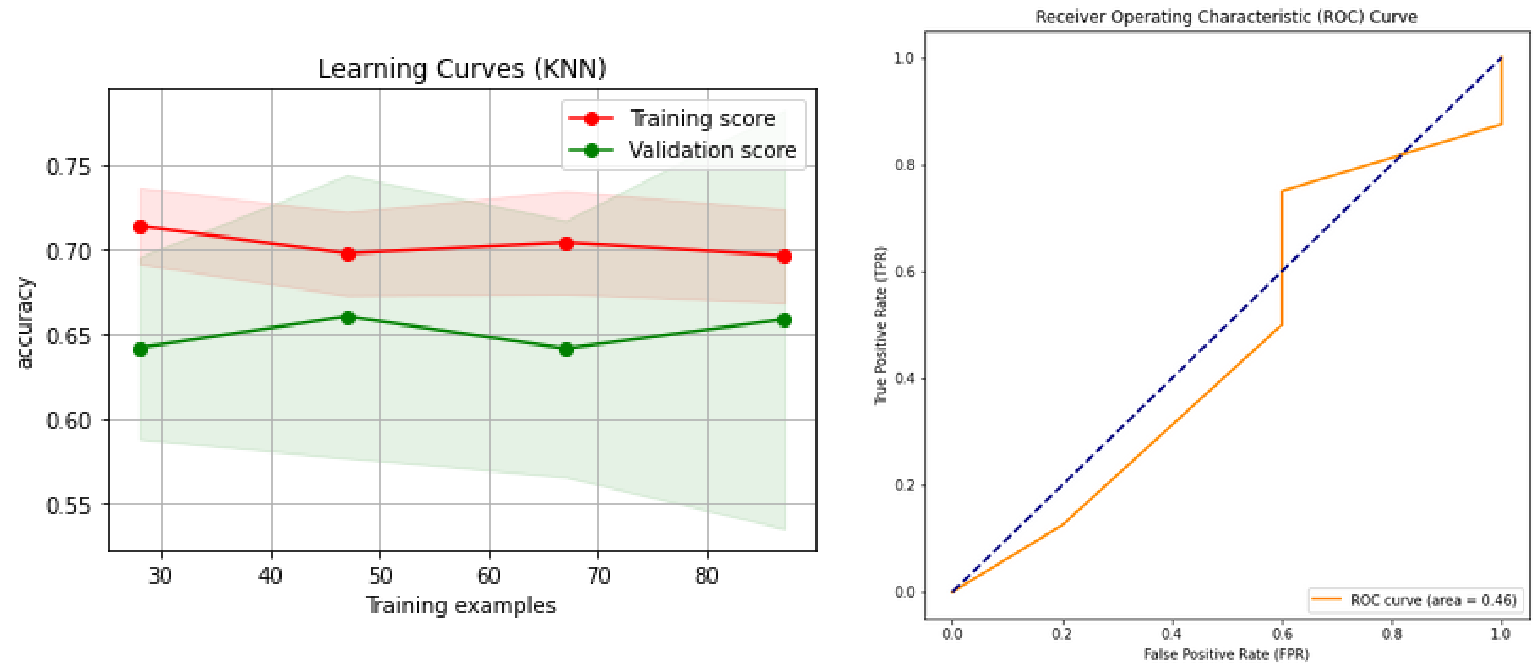
Feature Engineering Dataset

Accuracy: 0.62

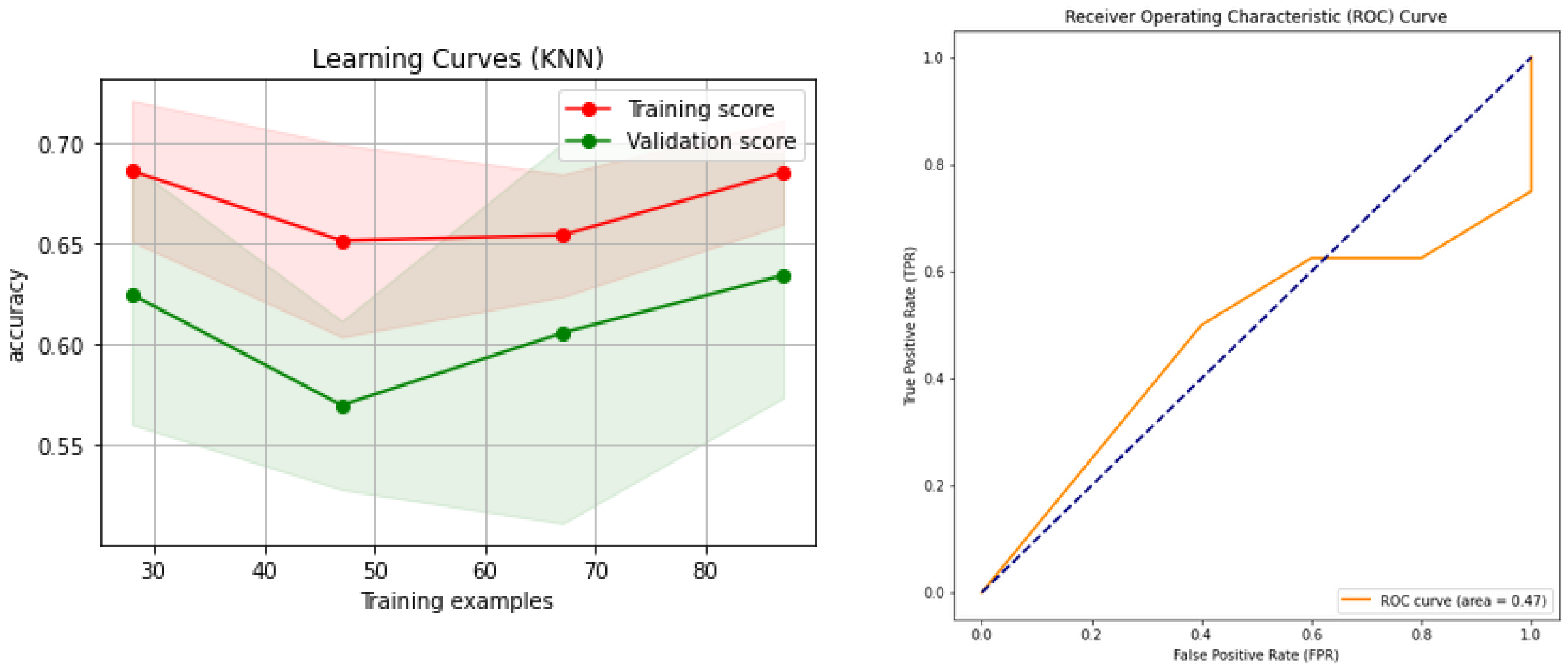
	precision	recall	f1-score	support
N	0.50	0.20	0.29	5
Y	0.64	0.88	0.74	8
accuracy			0.62	13
macro avg	0.57	0.54	0.51	13
weighted avg	0.58	0.62	0.56	13

Feature Engineering 2 Dataset
(year 9 prediction)

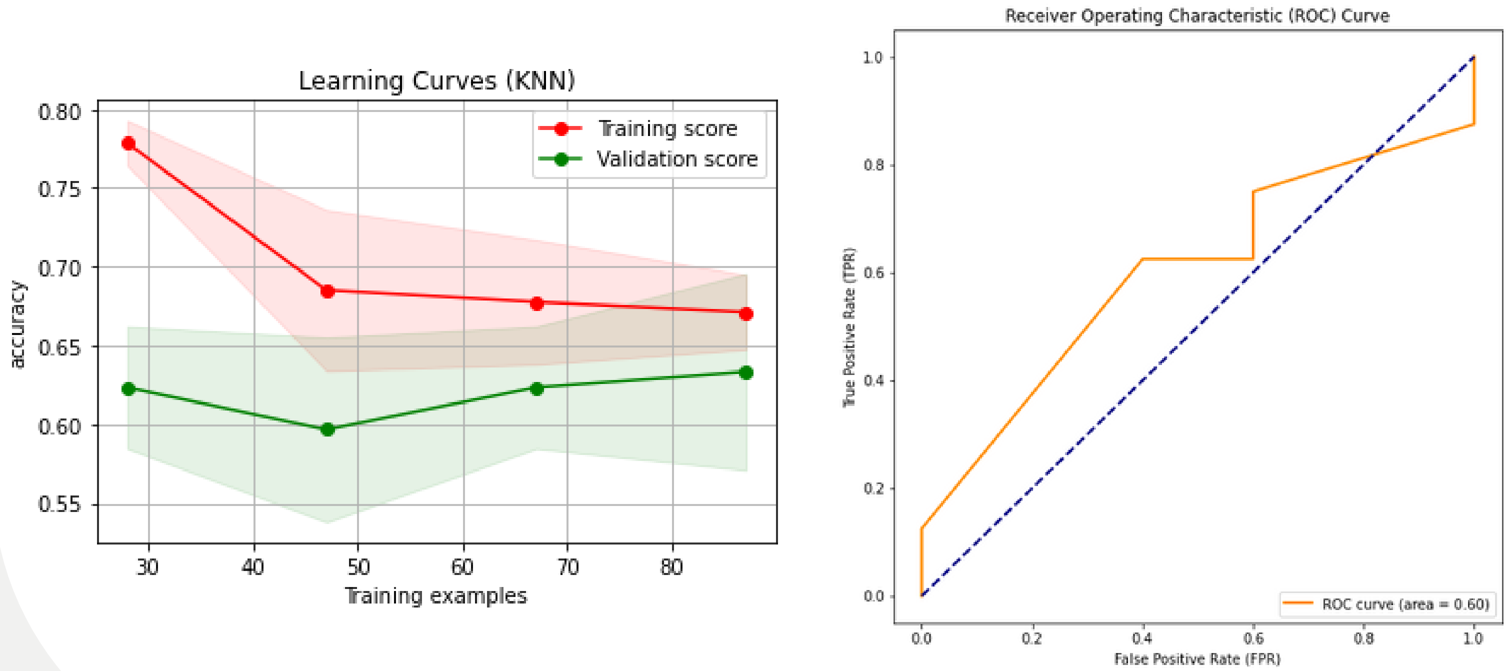
Annex - K Nearest Neighbors - Curves



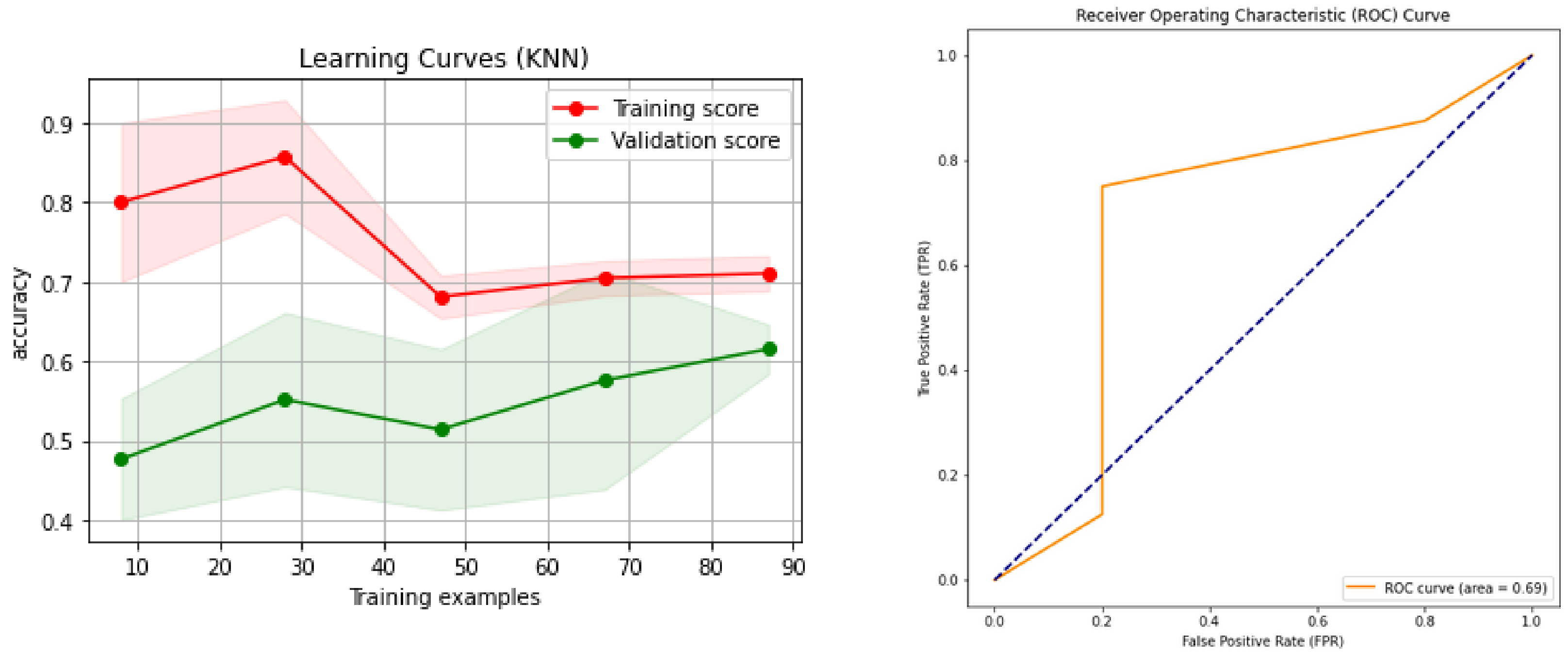
Original Dataset



Feature Engineering Dataset



Feature Selection Dataset



Feature Engineering 2 Dataset (year 9)

Annex - Decision Tree - Results

Accuracy: 0.38

	precision	recall	f1-score	support
0	0.38	1.00	0.56	5
1	1.00	0.00	0.00	8
accuracy			0.38	13
macro avg	0.69	0.50	0.28	13
weighted avg	0.76	0.38	0.21	13

Original Dataset

Accuracy: 0.38

	precision	recall	f1-score	support
0	0.38	1.00	0.56	5
1	1.00	0.00	0.00	8
accuracy			0.38	13
macro avg	0.69	0.50	0.28	13
weighted avg	0.76	0.38	0.21	13

Feature Selection Dataset

Accuracy: 0.38

	precision	recall	f1-score	support
0	0.38	1.00	0.56	5
1	1.00	0.00	0.00	8
accuracy			0.38	13
macro avg	0.69	0.50	0.28	13
weighted avg	0.76	0.38	0.21	13

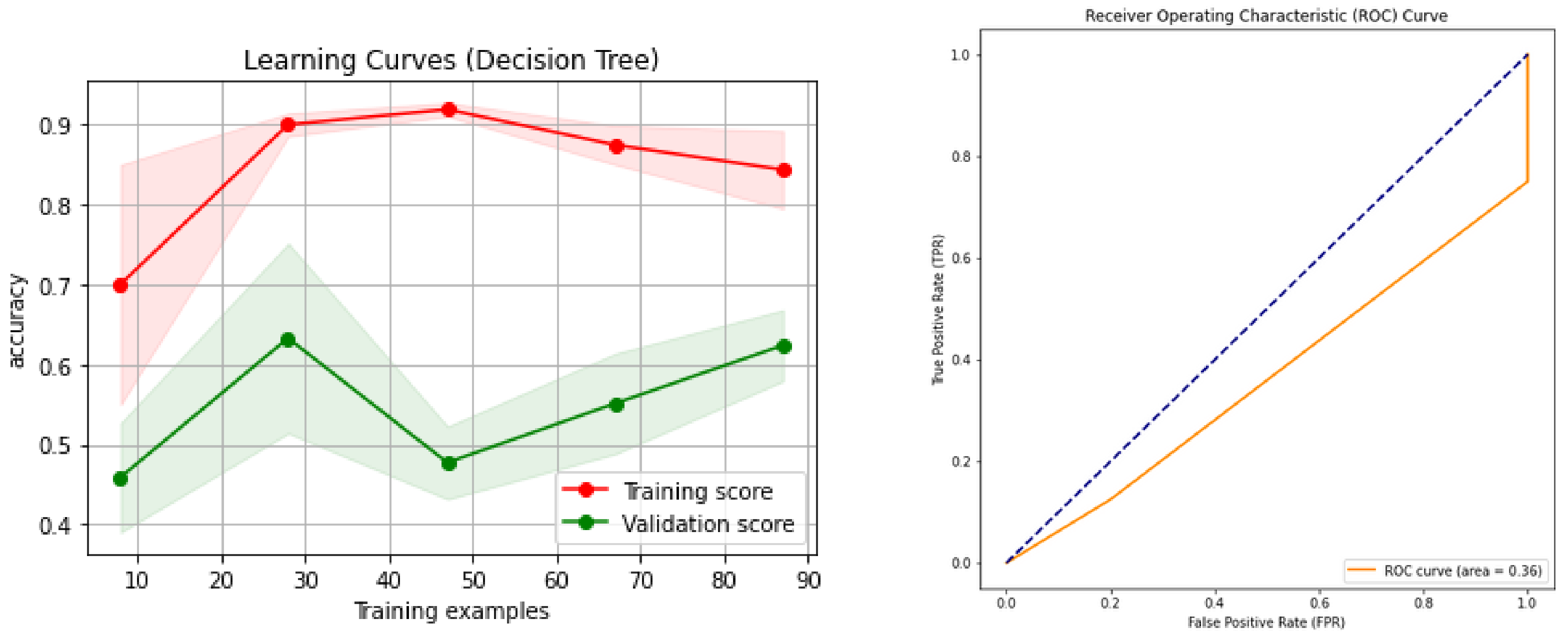
Feature Engineering Dataset

Accuracy: 0.38

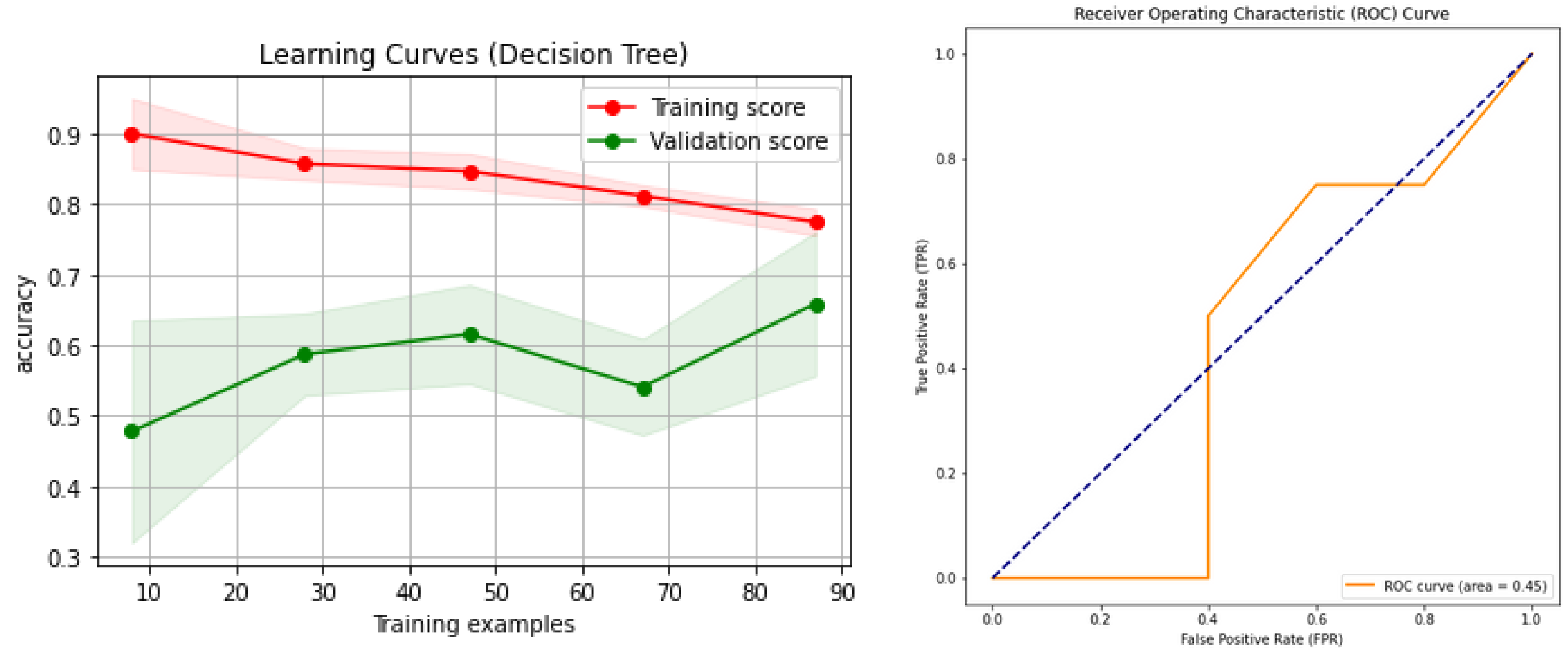
	precision	recall	f1-score	support
0	0.38	1.00	0.56	5
1	1.00	0.00	0.00	8
accuracy			0.38	13
macro avg	0.69	0.50	0.28	13
weighted avg	0.76	0.38	0.21	13

Feature Engineering 2 Dataset
(year 9 prediction)

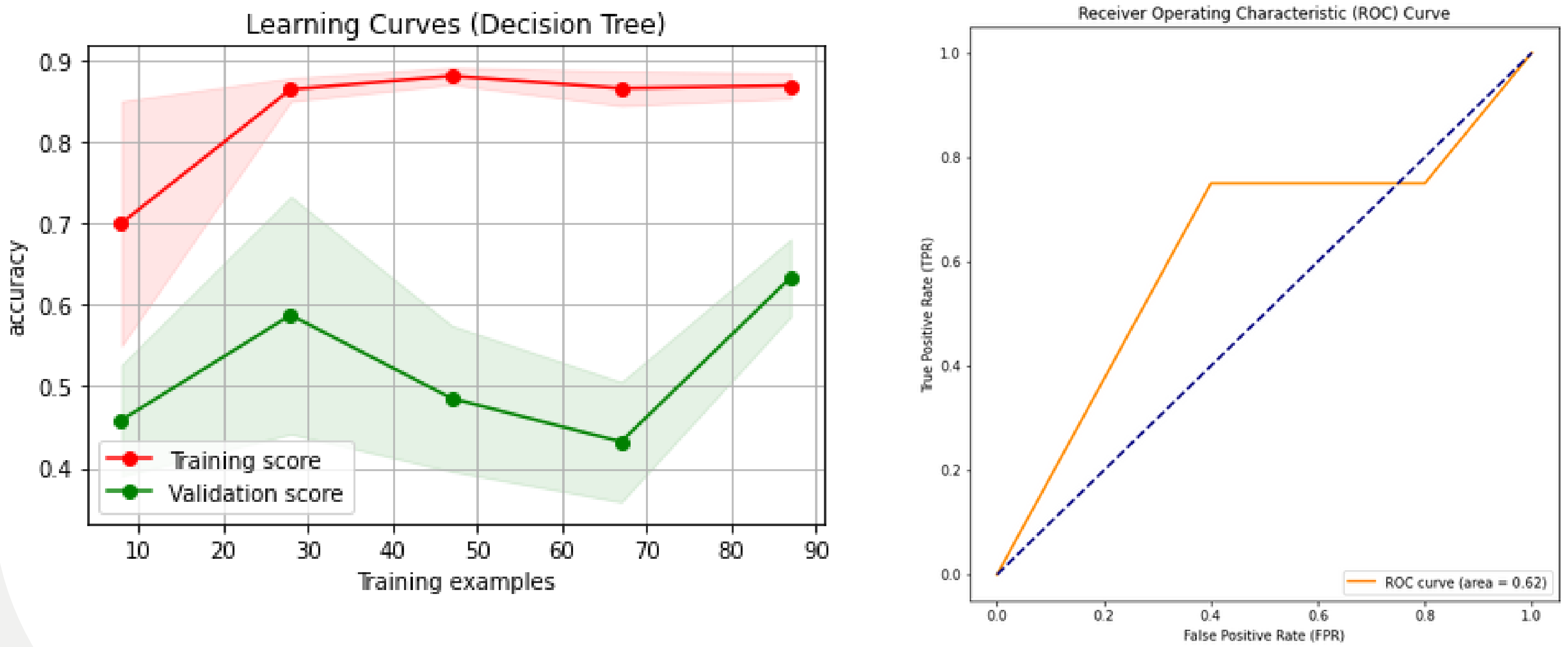
Annex - Decision Tree - Curves



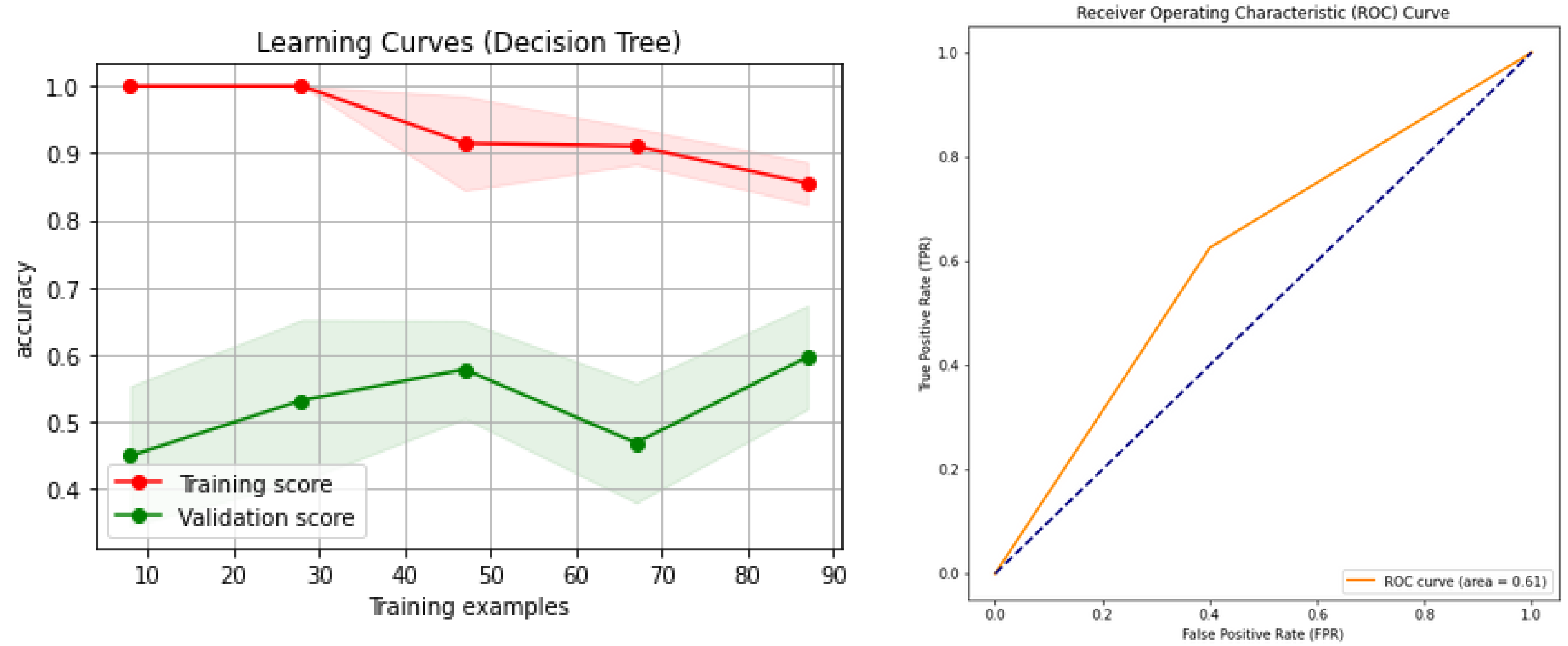
Original Dataset



Feature Engineering Dataset



Feature Selection Dataset



Feature Engineering 2 Dataset (year 9)

Annex - Random Forest- Results

Accuracy: 0.38

	precision	recall	f1-score	support
0	0.38	1.00	0.56	5
1	1.00	0.00	0.00	8
accuracy			0.38	13
macro avg	0.69	0.50	0.28	13
weighted avg	0.76	0.38	0.21	13

Original Dataset

Accuracy: 0.38

	precision	recall	f1-score	support
0	0.38	1.00	0.56	5
1	1.00	0.00	0.00	8
accuracy			0.38	13
macro avg	0.69	0.50	0.28	13
weighted avg	0.76	0.38	0.21	13

Feature Selection Dataset

Accuracy: 0.38

	precision	recall	f1-score	support
0	0.38	1.00	0.56	5
1	1.00	0.00	0.00	8
accuracy			0.38	13
macro avg	0.69	0.50	0.28	13
weighted avg	0.76	0.38	0.21	13

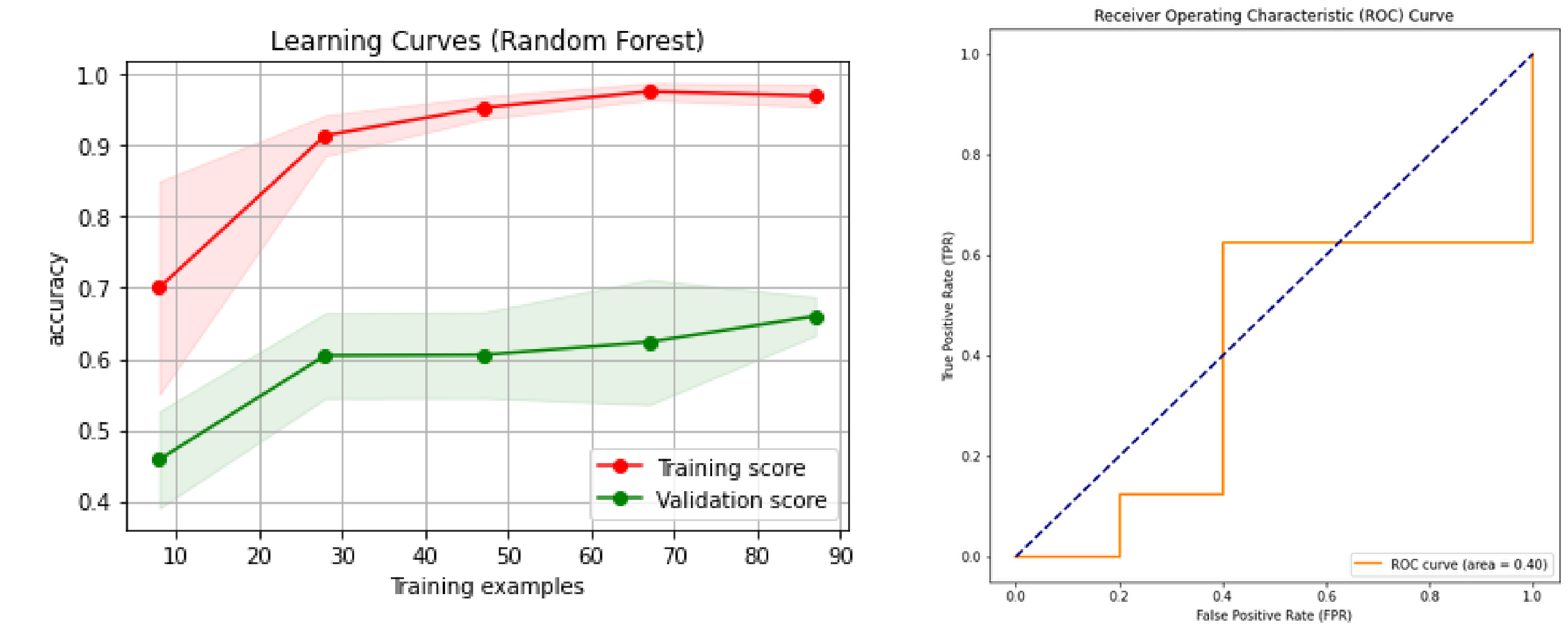
Feature Engineering Dataset

Accuracy: 0.38

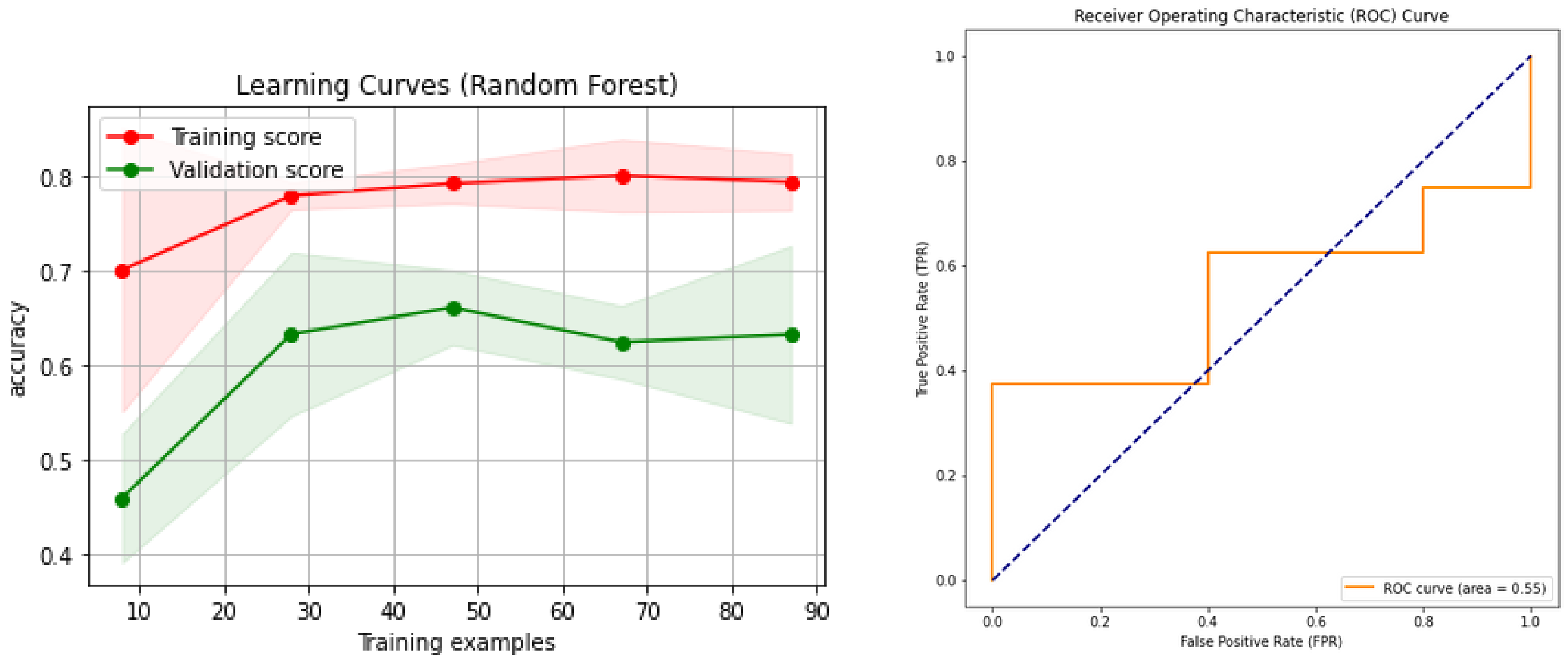
	precision	recall	f1-score	support
0	0.38	1.00	0.56	5
1	1.00	0.00	0.00	8
accuracy			0.38	13
macro avg	0.69	0.50	0.28	13
weighted avg	0.76	0.38	0.21	13

Feature Engineering 2 Dataset
(year 9 prediction)

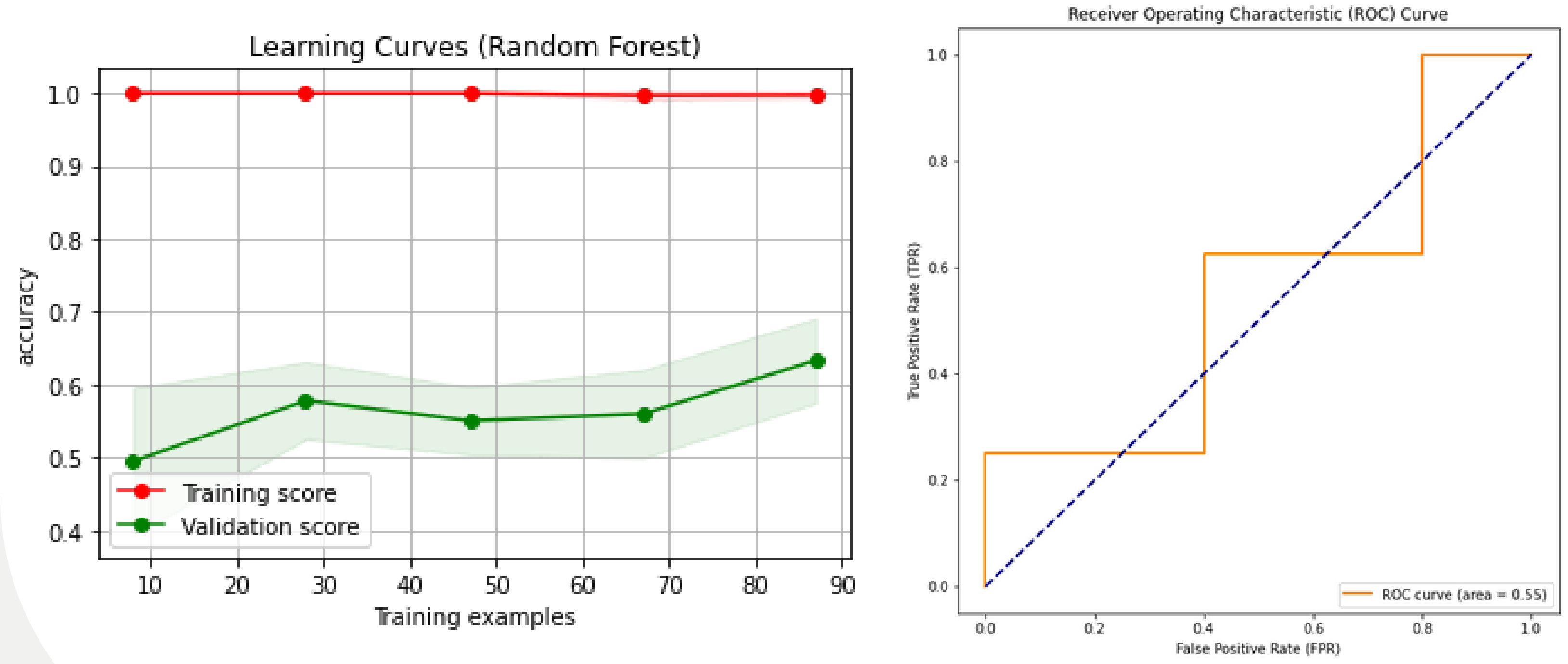
Annex - Random Forest - Curves



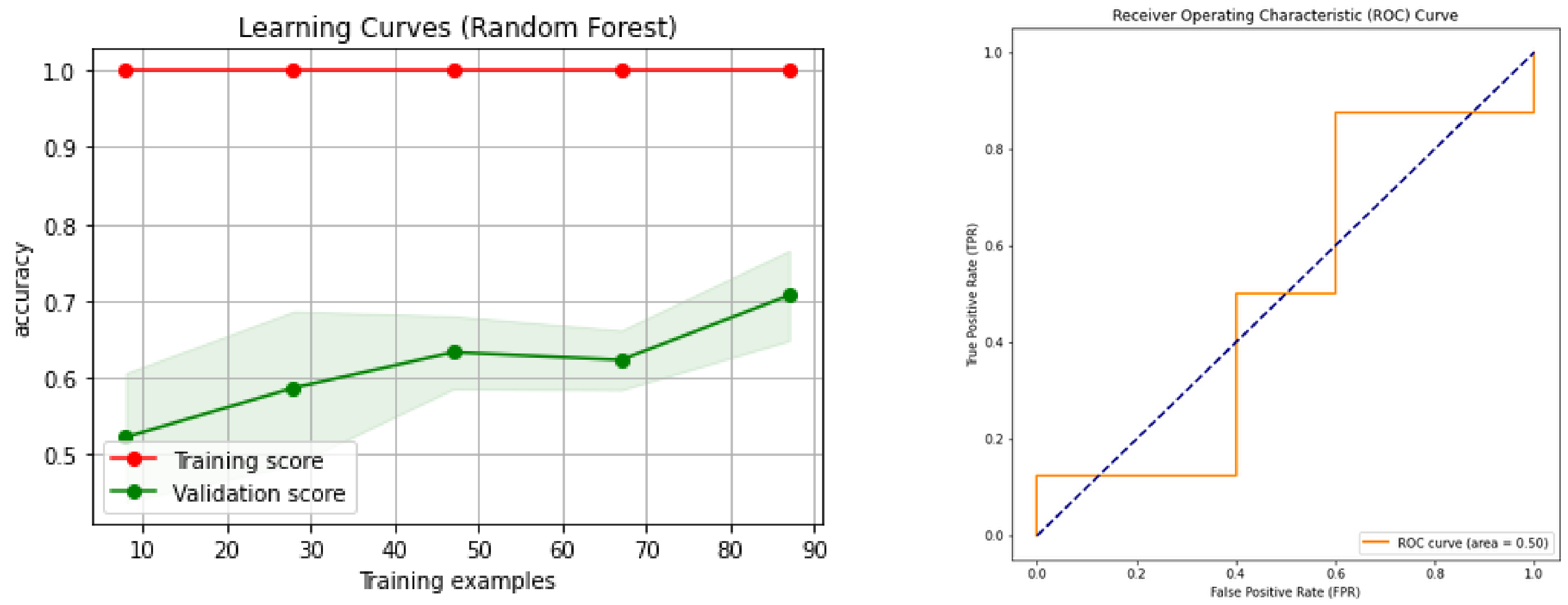
Original Dataset



Feature Engineering Dataset



Feature Selection Dataset



Feature Engineering 2 Dataset (year 9)