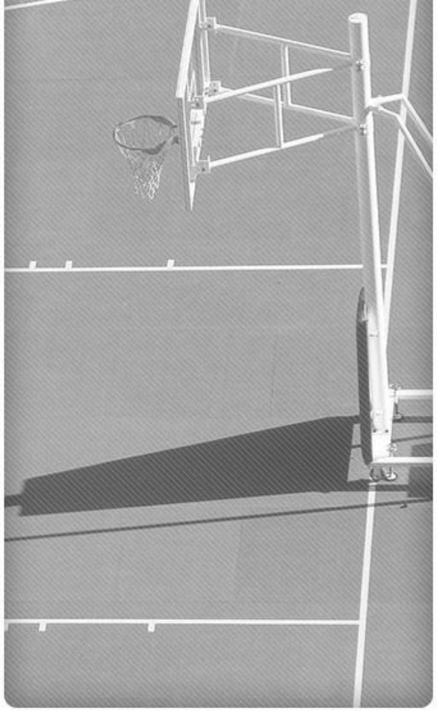


DATA MINING PROJECT

Aprendizagem Computacional
MEIC-001 2024/2025

Grupo 56
João Coutinho - up202108787
João Miranda - up2021
Miguel Garrido - up2021

CONTENTS

01 **Problem Specification**

02 **Data Description**

03 **Exploratory Data Analysis**

04 **Data Preparation**

05 **Results**

Domain Description

WNBA Structure:

- *Regular Season* - Each team plays a set number of games, competing to secure a spot in the playoffs.
- *Playoffs* - A bracket-style format is used to determine the league champion.

Dataset Overview:

- 10 years of data from players, teams, coaches, games, comprised of several different metrics, were collected and arranged on this dataset.

Main Goal:

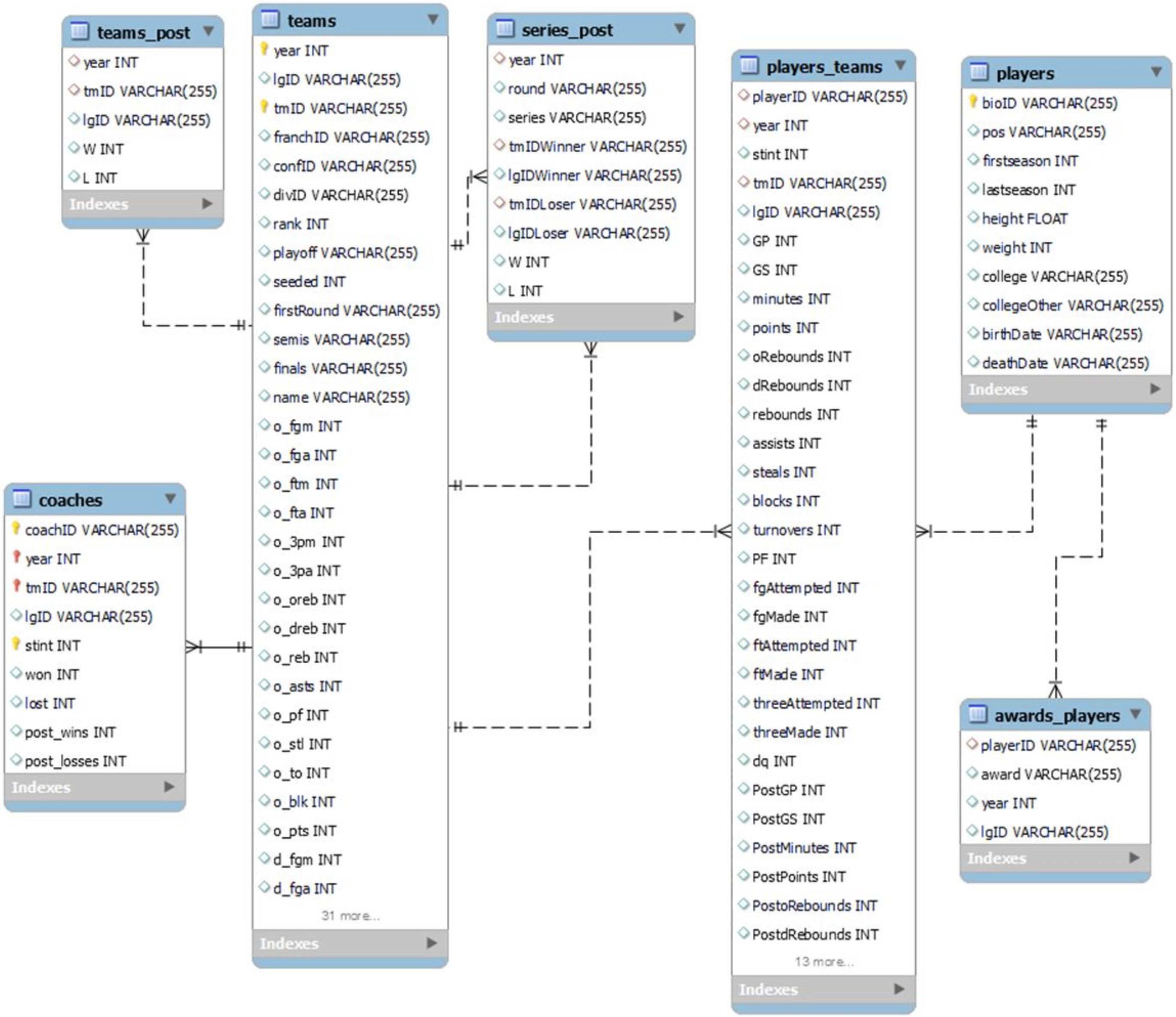
- The problem at hand requires the use of the available data to predict whether a team will qualify (or not) for the playoffs in the following season.

Domain Description

The provided data, in CSV format, contains the following tables:

- ***awards_players*** (96 objects) – Each record describes awards and prizes received by players across the 10 seasons.
- ***coaches*** (163 objects) – Each record describes all coaches who've managed the teams during the given period.
- ***players*** (894 objects) – Each record contains details of all players who played in the WNBA across this period.
- ***players_teams*** (1877 objects) – Each record describes the performance of each player for each team they played for.
- ***series_post*** (71 objects) – Each record describes the series' results.
- ***teams*** (143 objects) – Each record describes the seasonal performance of each team.
- ***teams_post*** (81 objects) – Each record describes the results of each team during the postseason.

Domain Description



teams_post
- year INT
- tmID VARCHAR(255)
- lgID VARCHAR(255)
- W INT
- L INT
- Indexes

teams
- year INT
- lgID VARCHAR(255)
- tmID VARCHAR(255)
- franchID VARCHAR(255)
- confID VARCHAR(255)
- divID VARCHAR(255)
- rank INT
- playoff VARCHAR(255)
- seeded INT
- firstRound VARCHAR(255)
- semis VARCHAR(255)
- finals VARCHAR(255)
- name VARCHAR(255)
- o_fgm INT
- o_fga INT
- o_ftm INT
- o_fta INT
- o_3pm INT
- o_3pa INT
- o_oreb INT
- o_dreb INT
- o_reb INT
- o_asts INT
- o_pf INT
- o_stl INT
- o_to INT
- o_blk INT
- o_pts INT
- d_fgm INT
- d_fga INT
- 31 more...
- Indexes

series_post
- year INT
- round VARCHAR(255)
- series VARCHAR(255)
- tmIDWinner VARCHAR(255)
- lgIDWinner VARCHAR(255)
- tmIDLoser VARCHAR(255)
- lgIDLoser VARCHAR(255)
- W INT
- L INT
- Indexes

players_teams
- playerID VARCHAR(255)
- year INT
- stint INT
- tmID VARCHAR(255)
- lgID VARCHAR(255)
- GP INT
- GS INT
- minutes INT
- points INT
- oRebounds INT
- dRebounds INT
- rebounds INT
- assists INT
- steals INT
- blocks INT
- turnovers INT
- PF INT
- fgAttempted INT
- fgMade INT
- ftAttempted INT
- ftMade INT
- threeAttempted INT
- threeMade INT
- dq INT
- PostGP INT
- PostGS INT
- PostMinutes INT
- PostPoints INT
- PostoRebounds INT
- PostdRebounds INT
- 13 more...
- Indexes

players
- bioID VARCHAR(255)
- pos VARCHAR(255)
- firstseason INT
- lastseason INT
- height FLOAT
- weight INT
- college VARCHAR(255)
- collegeOther VARCHAR(255)
- birthDate VARCHAR(255)
- deathDate VARCHAR(255)
- Indexes

coaches
- coachID VARCHAR(255)
- year INT
- tmID VARCHAR(255)
- lgID VARCHAR(255)
- stint INT
- won INT
- lost INT
- post_wins INT
- post_losses INT
- Indexes

awards_players
- playerID VARCHAR(255)
- award VARCHAR(255)
- year INT
- lgID VARCHAR(255)
- Indexes

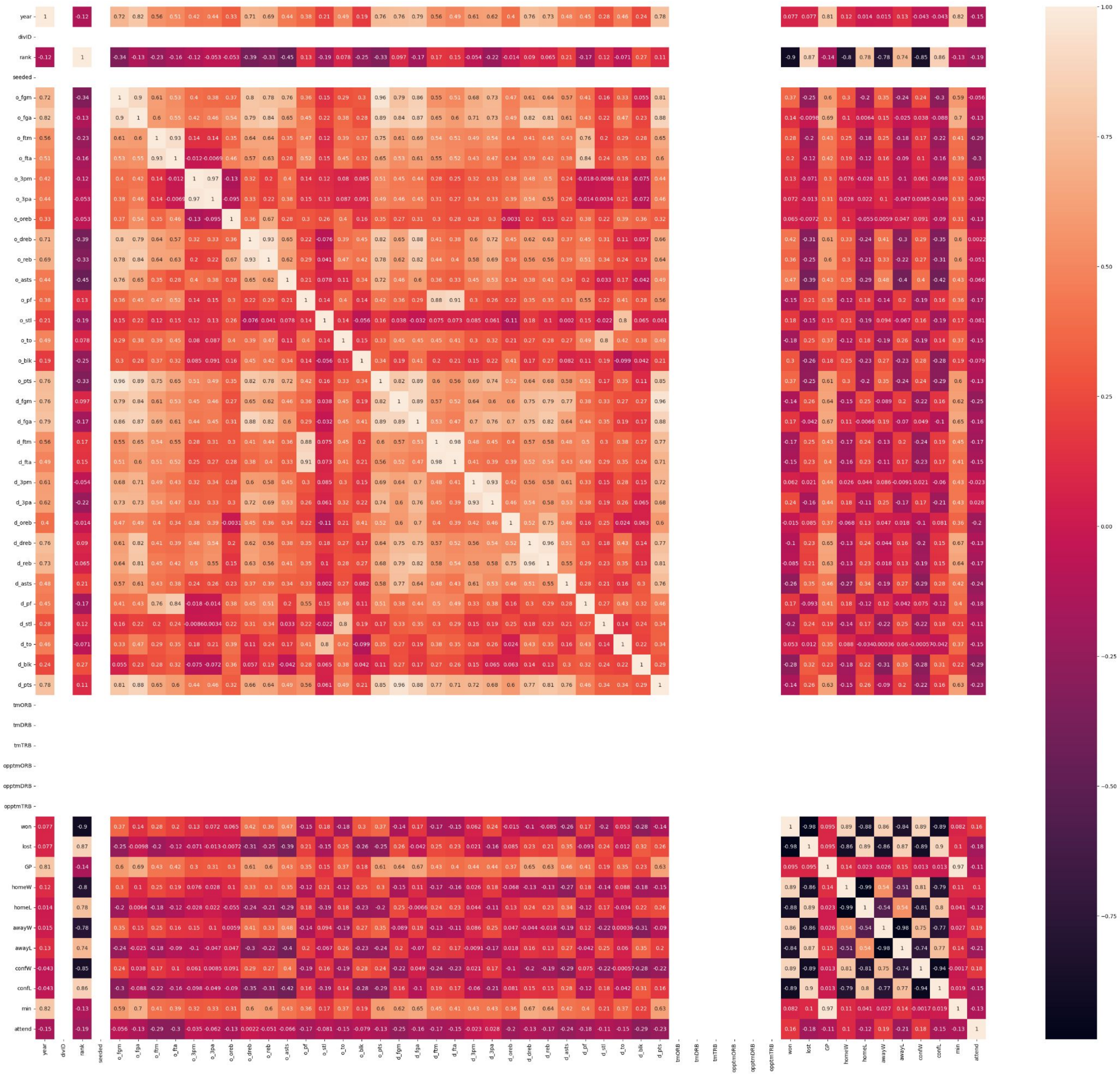
Exploratory Data Analysis

In order to explore correlations amongst various features, we generated several plots, focusing on identifying information that might be redundant as well as features that could be irrelevant or unnecessary to players/teams statistics and, thus, to the problem we had in hand.

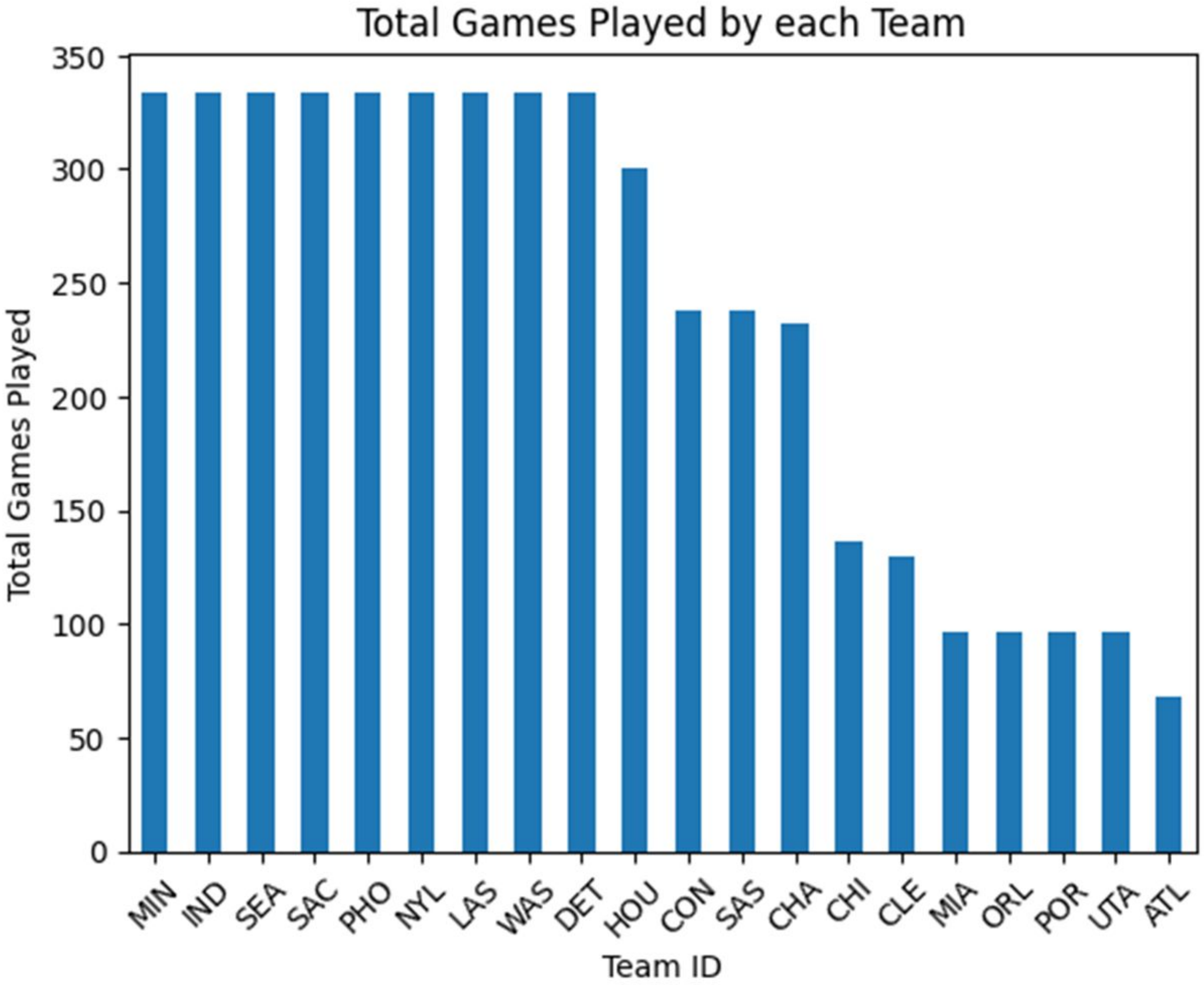
Besides a correlation matrix, other examples of our analysis are:

- Total games played by each team;
- Playoff appearances of each team;
- Box plots aiming to check the interquartile range of different player stats.

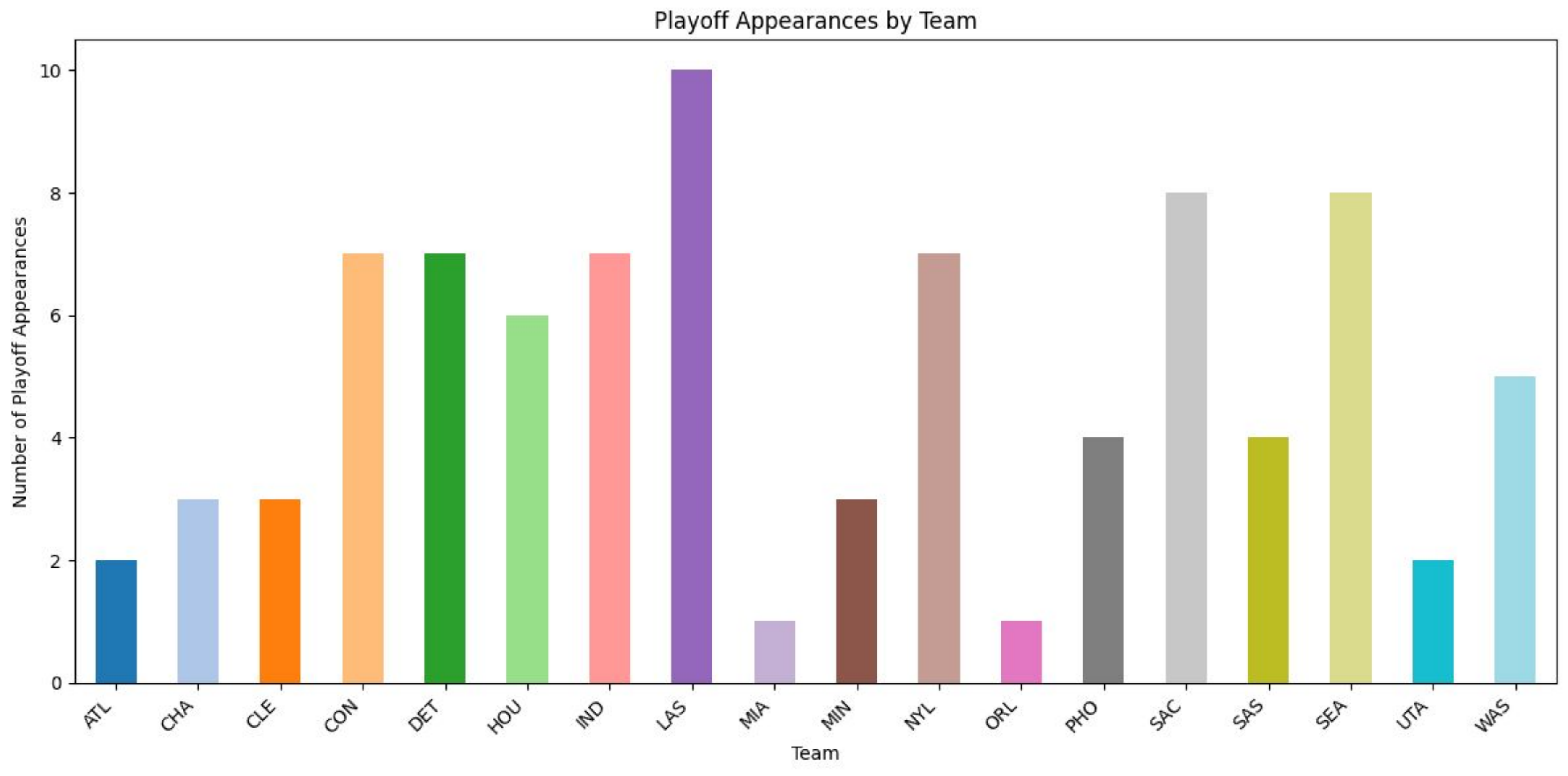
Exploratory Data Analysis



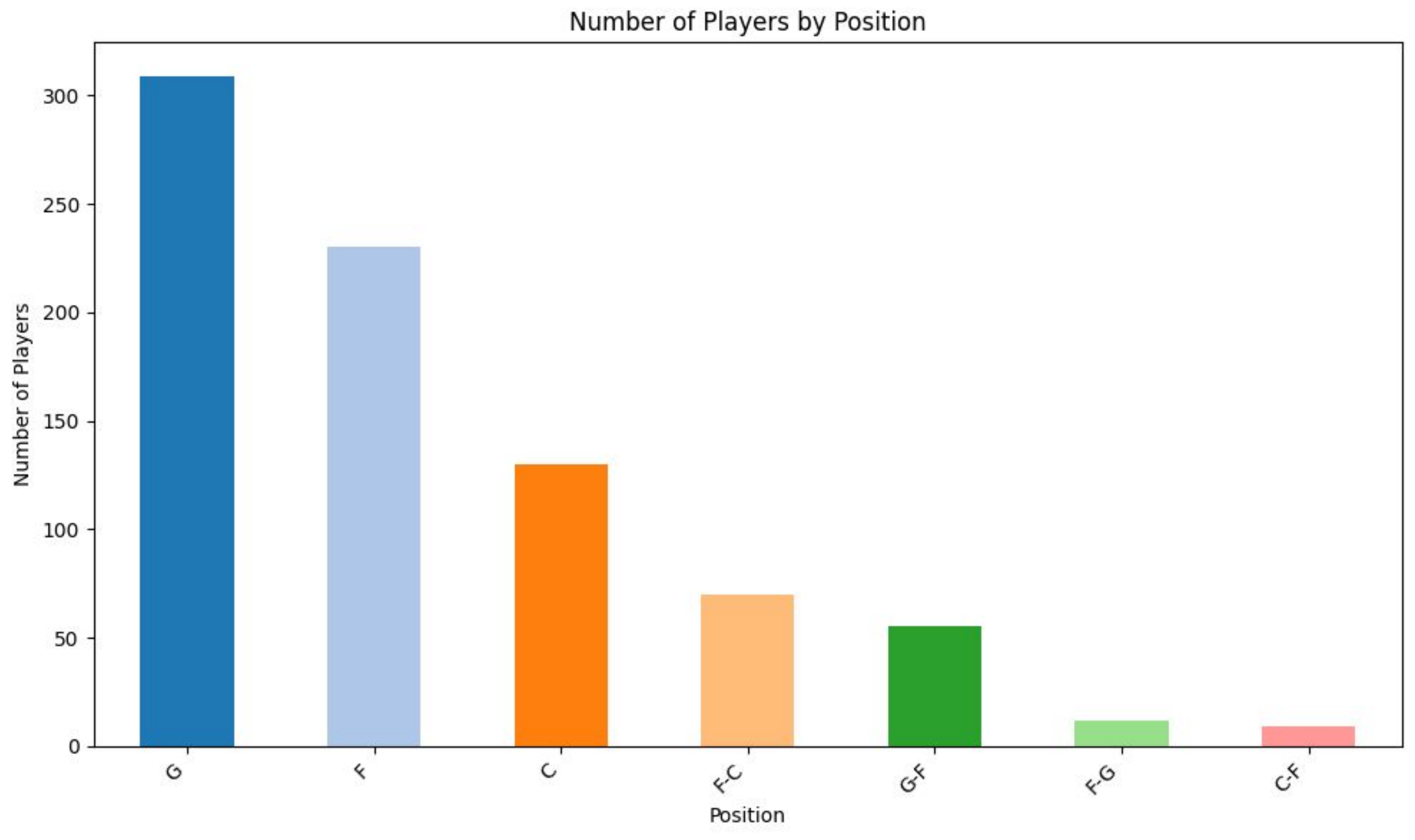
Exploratory Data Analysis



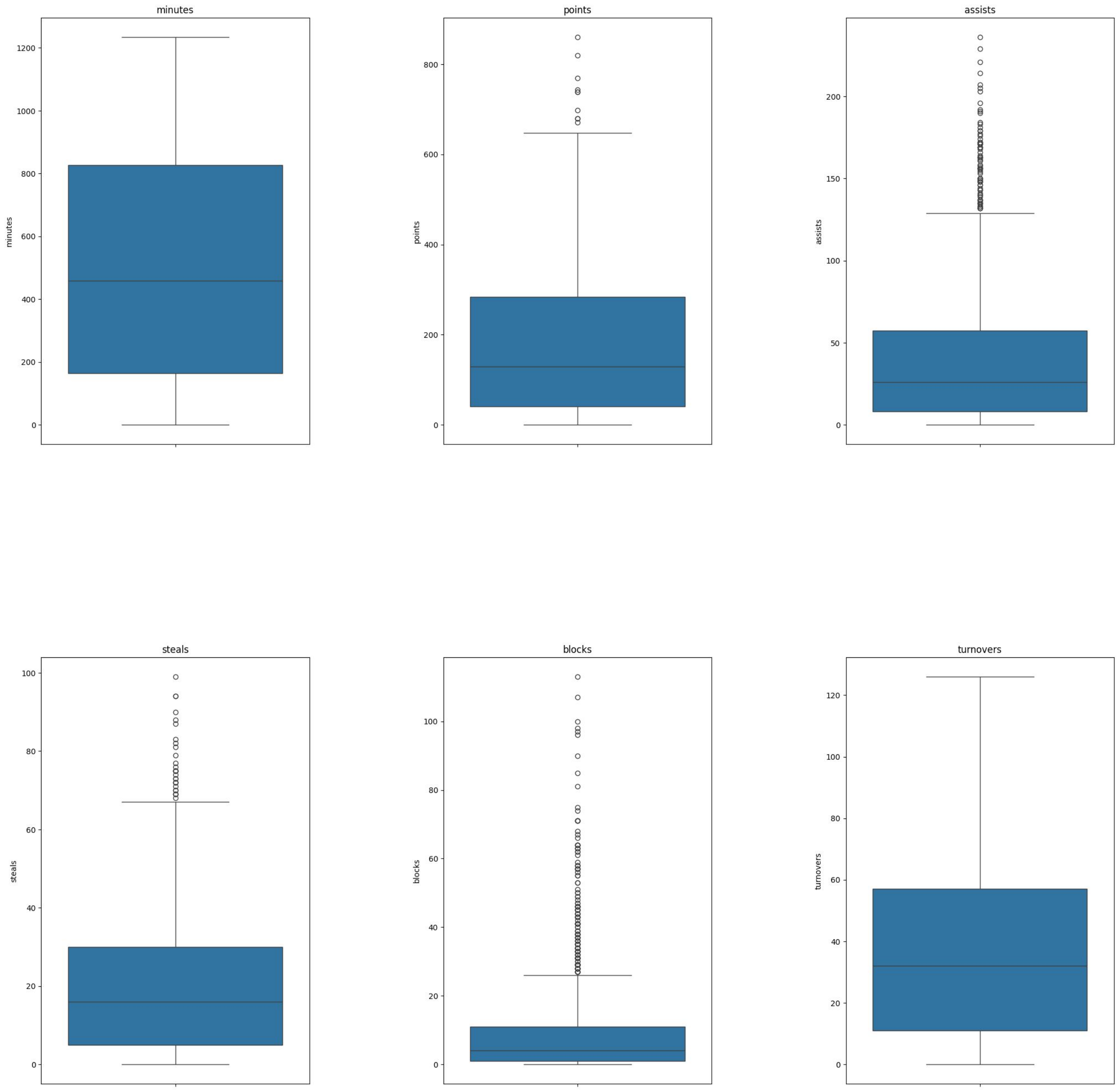
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Problem Definition

OBJECTIVE

The goal of this data mining project is to predict which teams in the Women's National Basketball Association (WNBA) will qualify for the playoffs in the next season based on team statistics, player performances and other data from previous years. To achieve this, 10 years of data are provided, allowing us to select and combine relevant features to create a final dataset. This dataset must then be properly cleaned and filtered before being tested and evaluated using metrics such as accuracy, precision, and others.

Data Preparation

FEATURE SELECTION

- ***teams*** – Replaced values that are empty strings with *NaN* and subsequently remove missing values. Irrelevant columns like *lgID*, *attend*, *seeded* or *arena* were also dropped.
- ***players*** – Dropped irrelevant columns like *college*, *birthDate* and *deathDate*, while also removing rows with impossible values for *height* and *weight* (less than zero) or with no assigned *pos* (field position) whatsoever.
- ***coaches*** – Removed irrelevant columns, like *lgID*. The *stint* column was also analyzed to make sure it was relevant for the data collection (the column had at least two different values).

For the remaining tables, essentially irrelevant attributes like *lgID* or columns with only a single unique value were dropped

Data Preparation

FEATURE ENGINEERING

For Players:

- Number of relevant awards, were considered relevant the following awards: "All-Star Game Most Valuable Player", "Defensive Player of the Year", "Most Improved Player", "Most Valuable Player", "Rookie of the Year" and "WNBA Finals Most Valuable Player"
- *Cumulative PER* (Player Efficiency Rating)

For Coaches:

- *Cumulative Win Rate* column, which represents a coach's career win rate (ranging from 0 to 1). For rookie coaches, where this value would be *NaN*, we replaced it with the average *Cumulative Win Rate* to give them the benefit of the doubt.
- Number of “Coach of the Year” awards

Data Preparation

FEATURE ENGINEERING

For Teams:

- *Cumulative PER* - the average of the score of the top 5 players on each team. These players must have logged more than 100 minutes combined across the regular and postseason in the previous season. The score is calculated by summing the *Cumulative PER* to the number of trophies * 0.005

Experimental Setup

TRAINING FEATURES

Besides *Score* from ***teams*** table and *Cumulative Win Rate* from ***coaches*** table, two features already provided in the dataset were used:

- *Rank*:
  - A team’s rank provides a direct measure of its performance relative to other teams in the league. We consider that teams with higher ranks in the previous year are more likely to qualify.
- *Won*:
  - Representing the total number of games won by the team during the previous season, it can demonstrate how a team previously behave in the regular season.

Experimental Setup

HYPERPARAMETER TUNING

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

CUSTOM BINARY CLASSIFICATION

Binary classification for *playoffNextYear* (0, 1).

SELECTING TOP 4 TEAMS OF EACH CONFERENCE

To guarantee exactly 8 teams pass to the playoffs, the teams with the 8 highest probabilities (4 for each conference) were chosen.

Results

Before the competition took place, we selected our top 3 classifiers based on the precision results obtained across all years; before the final delivery, we added yet another classifier to the selection pool:

DECISION TREES (DT)

- tree-like structures to make decisions
- intuitive and easy to interpret, built-in feature selection, doesn’t require data normalization
- limited predictive power, prone to overfitting

SUPPORT VECTOR MACHINES (SVM)

- find a hyperplane in an N-dimensional space that separates and classifies data
- good resistance to overfitting, effective in high-dimensional spaces (datasets with many features)
- sensitive to both hyperparameter tuning and noise, computationally expensive, not advisable for large datasets

Results

ADABOOST (AB)

- creates a strong classifier by combining multiple weaker classifiers (by fitting them in constantly modified versions of the dataset)
- high level of precision, resistant to overfitting, versatile
- sensitive to noise and outliers, resource intensive

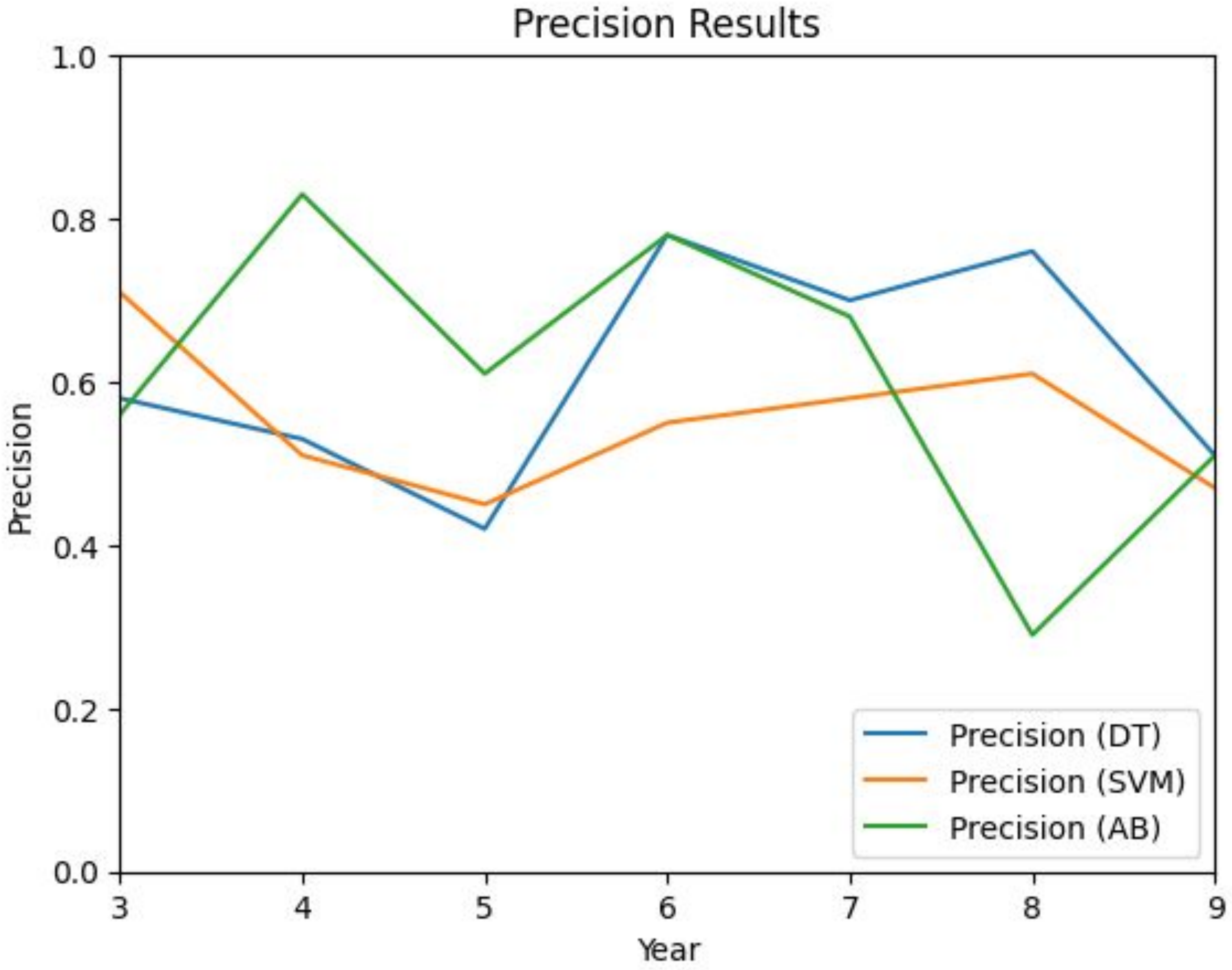
K-NEAREST NEIGHBORS (KNN)

- classifies data points based on the majority class of their nearest neighbors in the feature space
- no assumptions about data distribution
- sensitive to irrelevant features and noisy data, requires careful selection of the number of neighbors (K)

We then analyzed the precision results for all three classifiers to determine which model performed best and was most suitable to achieve the best result possible in the competition.

Results (Competition Version)

- Decision Trees seemed to perform better overall in the latter years
- AdaBoost doesn’t work with the set of probabilities we required



Competition Results (Day 1)

tmID	confID	predict
ATL	EA	0.53
CHI	EA	0.23
CON	EA	0.23
IND	EA	1.00
NYL	EA	0.53
WAS	EA	0.53
LAS	WE	1.00
MIN	WE	0.53
PHO	WE	1.00
SAS	WE	1.00
SEA	WE	0.53
TUL	WE	0.53

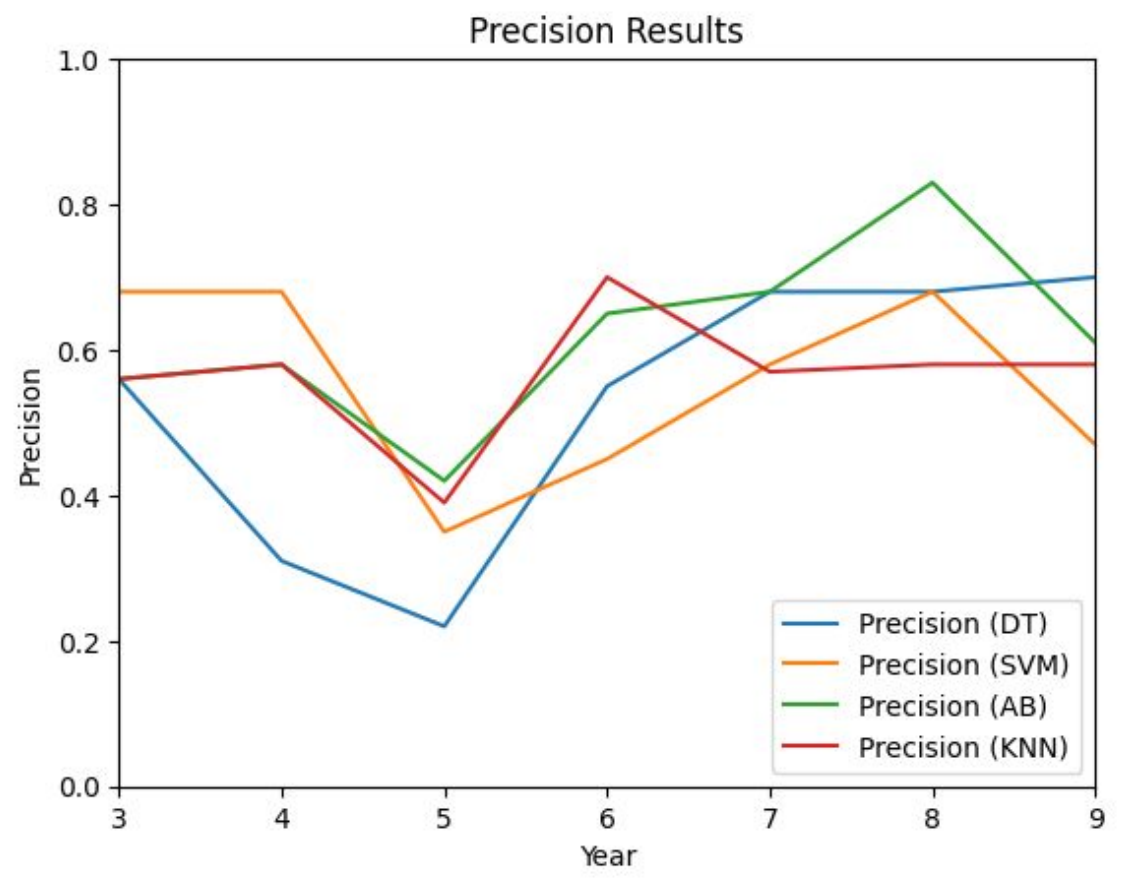
Competition Results (Day 2)

tmID	confID	predict
ATL	EA	0.57
CHI	EA	0.00
CON	EA	0.00
IND	EA	1.00
NYL	EA	0.86
WAS	EA	0.57
LAS	WE	0.78
MIN	WE	0.00
PHO	WE	1.00
SAS	WE	0.78
SEA	WE	0.86
TUL	WE	0.58

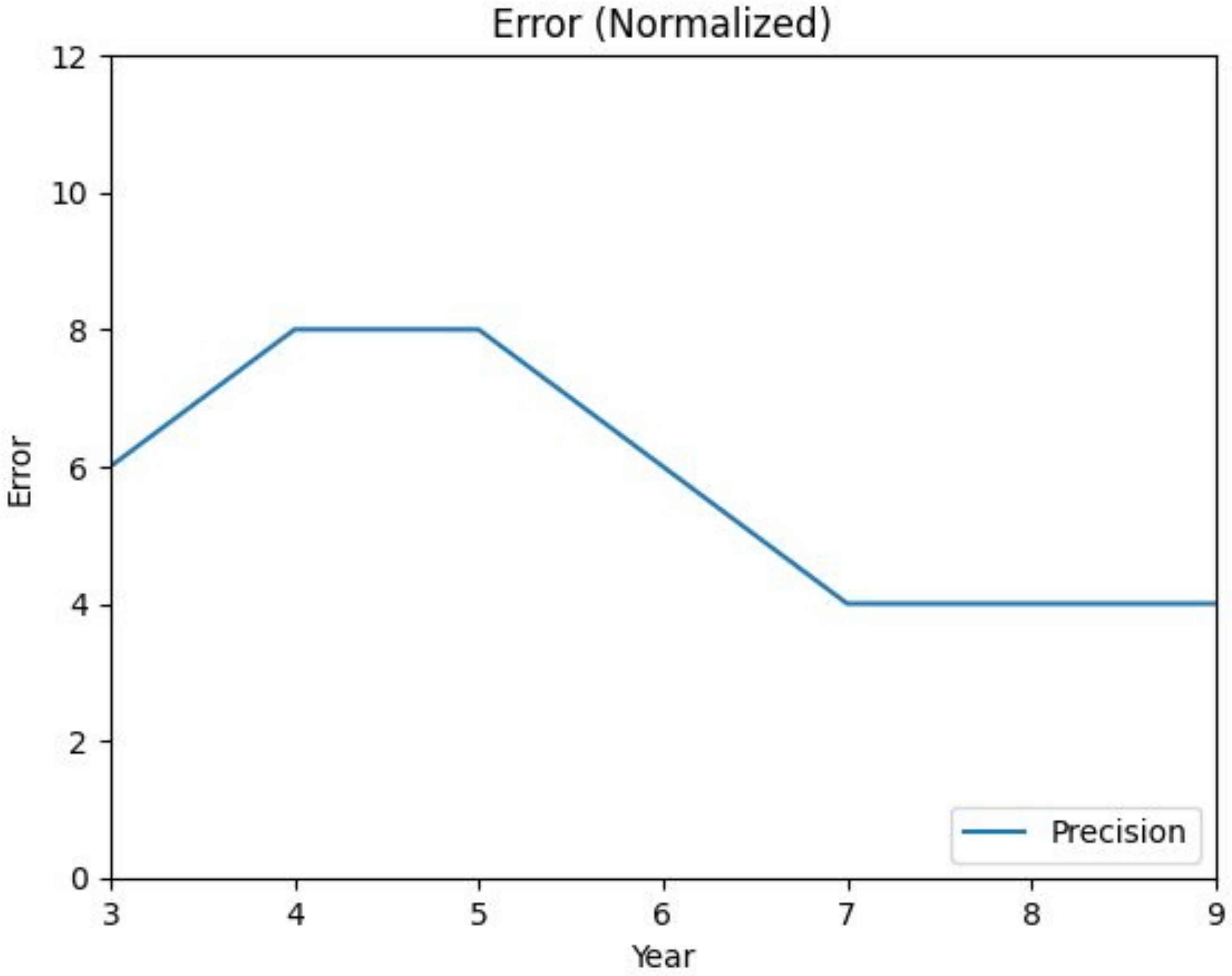
Results (Final Version)

IMPROVEMENTS

- Application of a cumulative PER instead of PER
- Score of each player and coach based also on respective awards
- Tests with the KNN classifier



Results (Final Version)



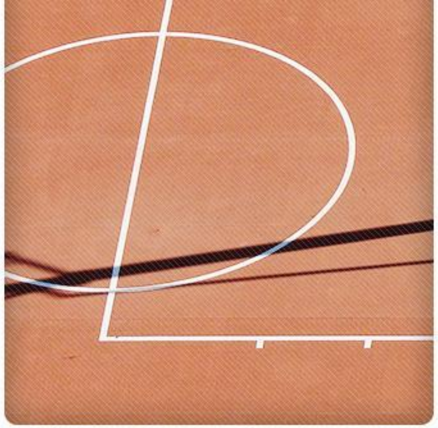
Final Results

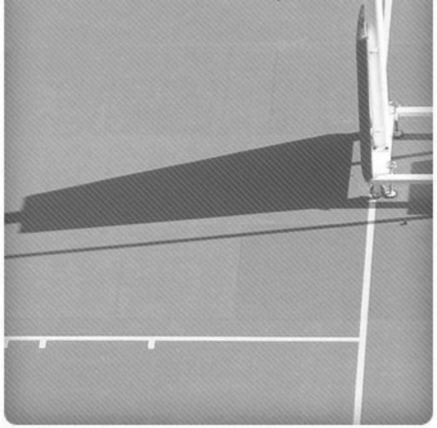
tmID	confID	predict	predictNormalized
ATL	EA	0.71	1
CHI	EA	0.43	0
CON	EA	1.00	1
IND	EA	1.00	1
NYL	EA	1.00	1
WAS	EA	0.43	0
LAS	WE	0.71	1
MIN	WE	0.43	0
PHO	WE	1.00	1
SAS	WE	1.00	1
SEA	WE	0.86	1
TUL	WE	0.12	0

Conclusions, limitations and future work

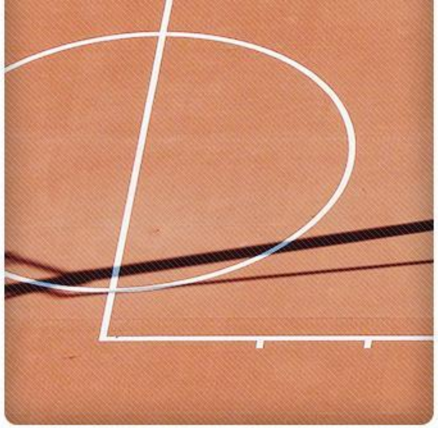
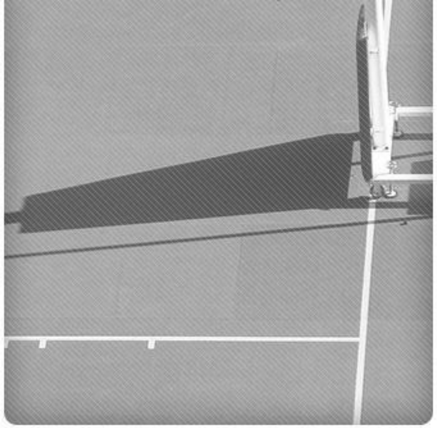
After completing this work, we feel we have gained essential insights in the process of machine learning and data mining technologies, as well as their practical application in forecasting and predictive analysis.

We ended up developing what we think is a reliable machine learning model to predict the WNBA playoff qualifiers with success, adopting an iterative approach with consistent and relevant improvements throughout every stage of the work.





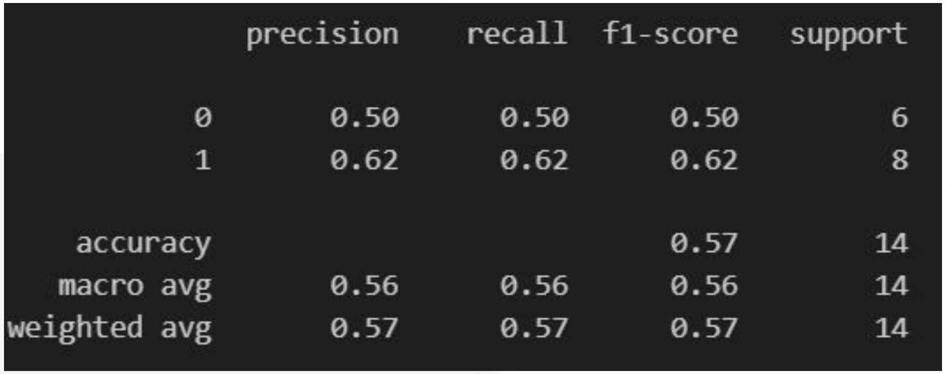
THANK YOU

ANNEXES

YEAR 3

DT

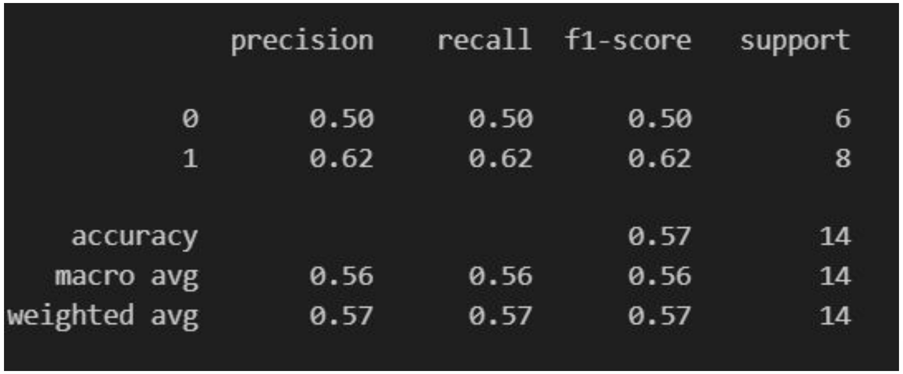


```
              precision    recall  f1-score   support

           0       0.50      0.50      0.50         6
           1       0.62      0.62      0.62         8

    accuracy                           0.57        14
   macro avg       0.56      0.56      0.56        14
weighted avg       0.57      0.57      0.57        14
```

AB

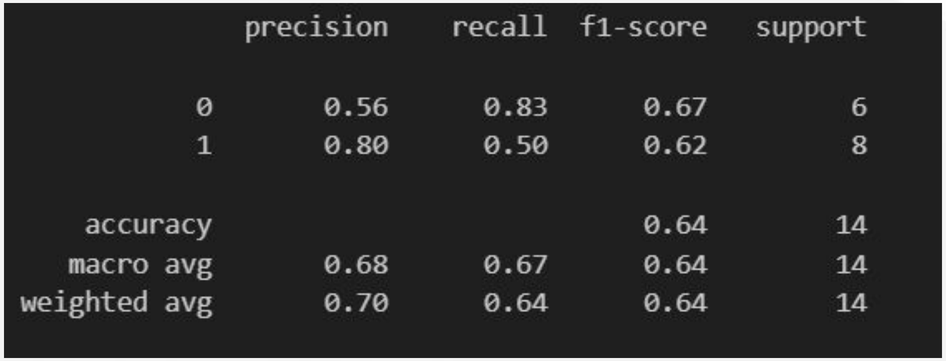


```
              precision    recall  f1-score   support

           0       0.50      0.50      0.50         6
           1       0.62      0.62      0.62         8

    accuracy                           0.57        14
   macro avg       0.56      0.56      0.56        14
weighted avg       0.57      0.57      0.57        14
```

SVM

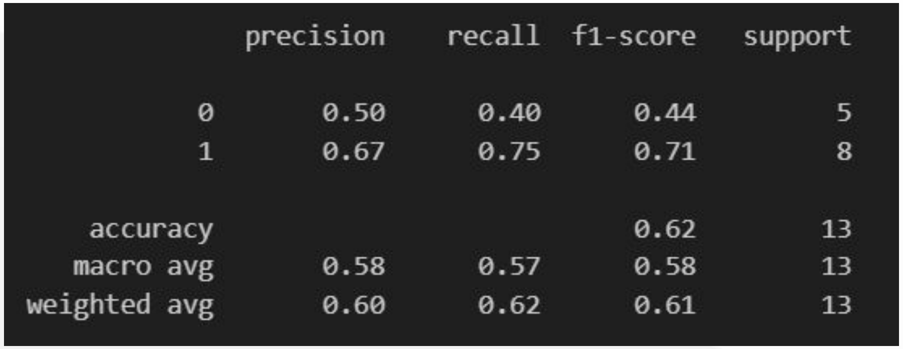


```
              precision    recall  f1-score   support

           0       0.56      0.83      0.67         6
           1       0.80      0.50      0.62         8

    accuracy                           0.64        14
   macro avg       0.68      0.67      0.64        14
weighted avg       0.70      0.64      0.64        14
```

KNN



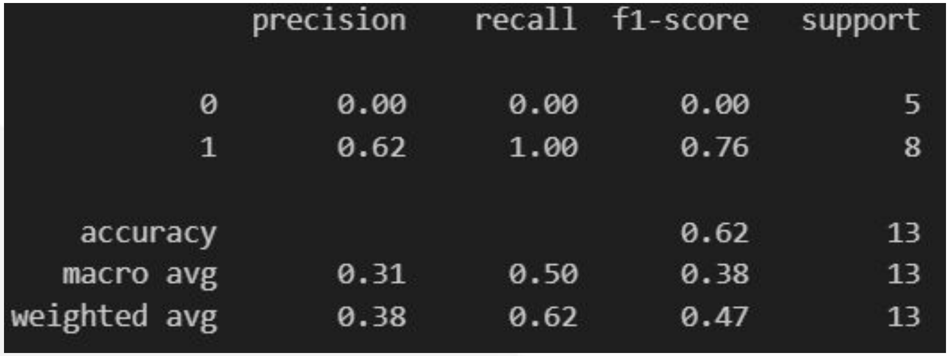
```
              precision    recall  f1-score   support

           0       0.50      0.40      0.44         5
           1       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
```

YEAR 4

DT

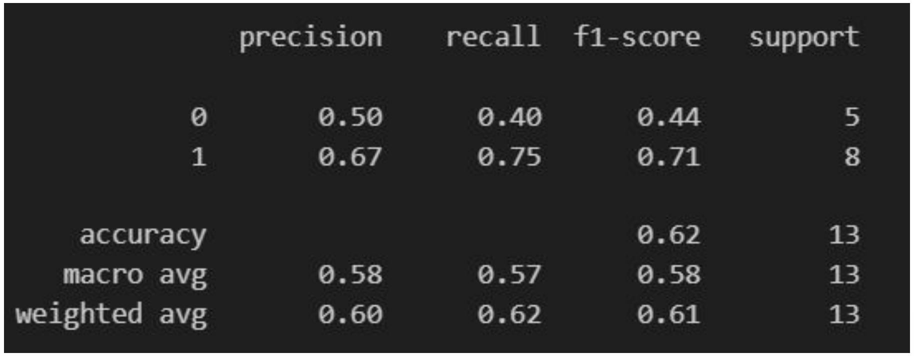


```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         5
           1       0.62      1.00      0.76         8

    accuracy                           0.62        13
   macro avg       0.31      0.50      0.38        13
weighted avg       0.38      0.62      0.47        13
```

AB

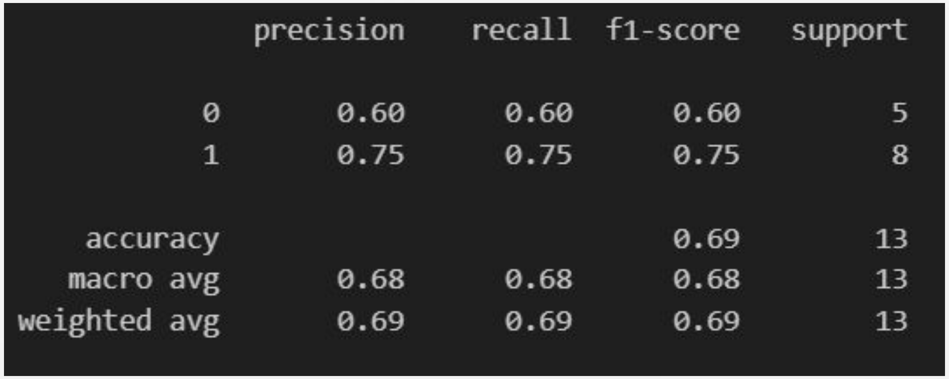


```
              precision    recall  f1-score   support

           0       0.50      0.40      0.44         5
           1       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
```

SVM

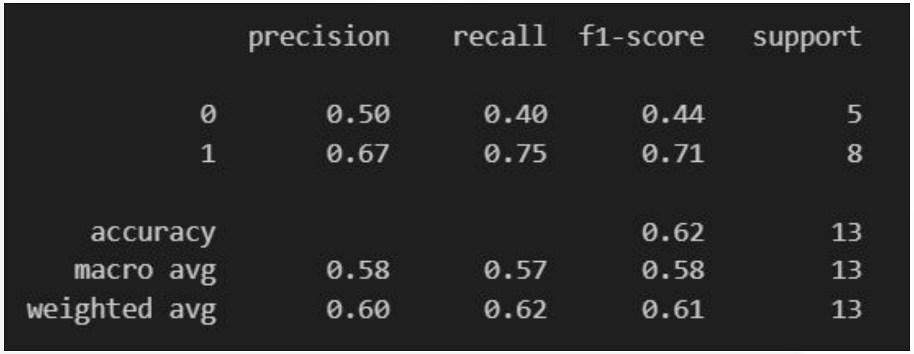


```
              precision    recall  f1-score   support

           0       0.60      0.60      0.60         5
           1       0.75      0.75      0.75         8

    accuracy                           0.69        13
   macro avg       0.68      0.68      0.68        13
weighted avg       0.69      0.69      0.69        13
```

KNN



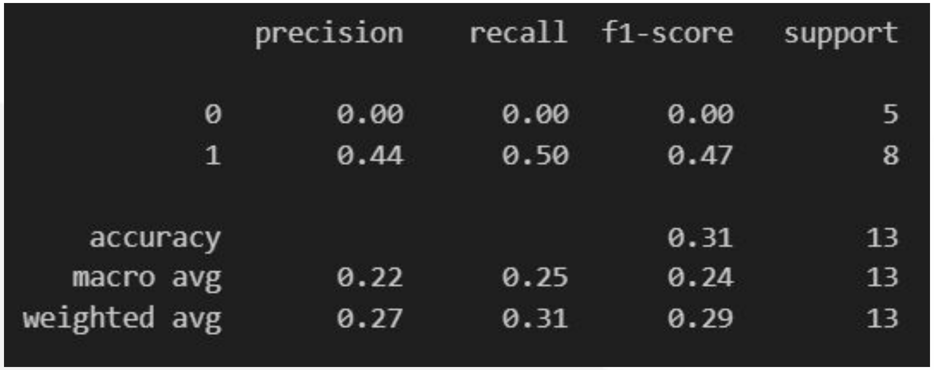
```
              precision    recall  f1-score   support

           0       0.50      0.40      0.44         5
           1       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
```

YEAR 5

DT

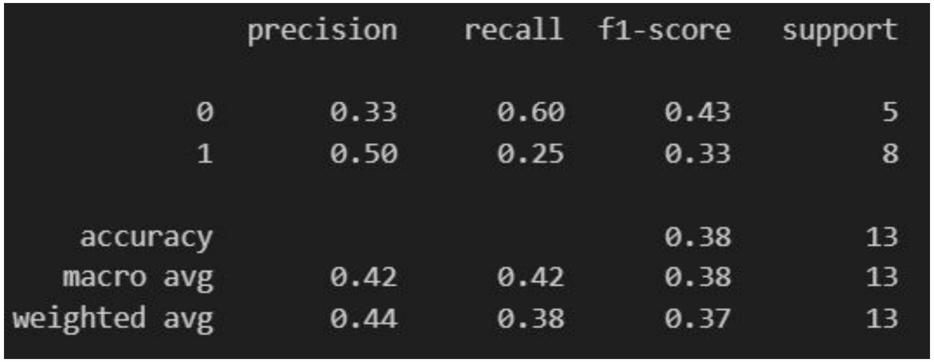


```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         5
           1       0.44      0.50      0.47         8

    accuracy                           0.31        13
   macro avg       0.22      0.25      0.24        13
weighted avg       0.27      0.31      0.29        13
```

AB

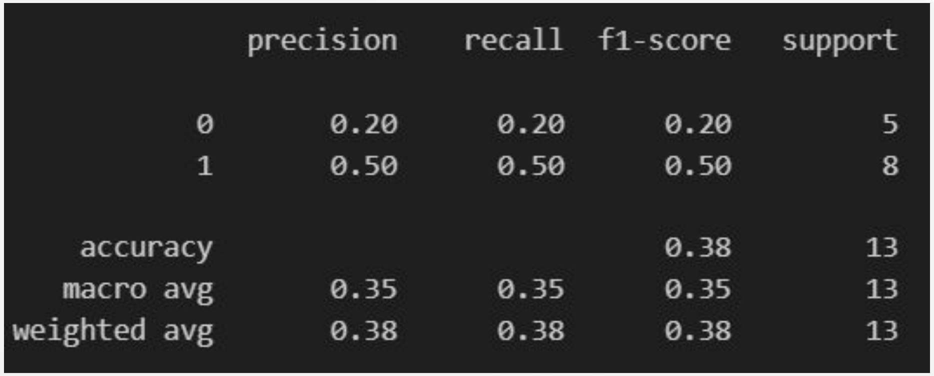


```
              precision    recall  f1-score   support

           0       0.33      0.60      0.43         5
           1       0.50      0.25      0.33         8

    accuracy                           0.38        13
   macro avg       0.42      0.42      0.38        13
weighted avg       0.44      0.38      0.37        13
```

SVM

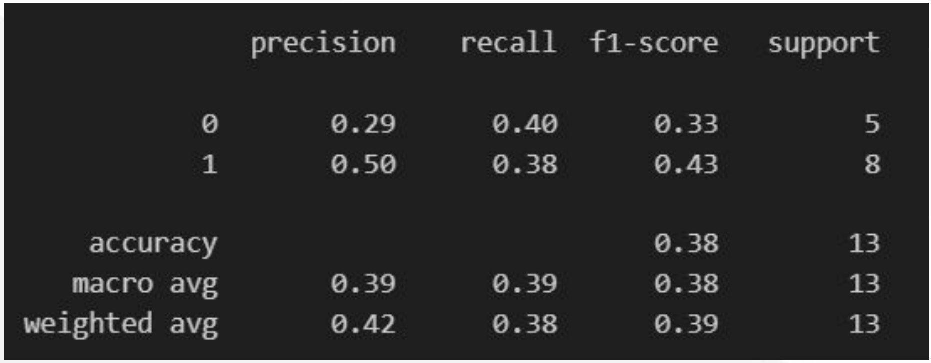


```
              precision    recall  f1-score   support

           0       0.20      0.20      0.20         5
           1       0.50      0.50      0.50         8

    accuracy                           0.38        13
   macro avg       0.35      0.35      0.35        13
weighted avg       0.38      0.38      0.38        13
```

KNN



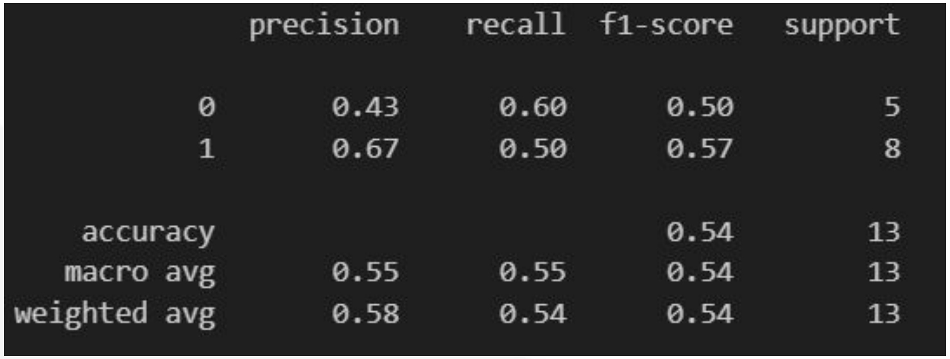
```
              precision    recall  f1-score   support

           0       0.29      0.40      0.33         5
           1       0.50      0.38      0.43         8

    accuracy                           0.38        13
   macro avg       0.39      0.39      0.38        13
weighted avg       0.42      0.38      0.39        13
```

YEAR 6

DT

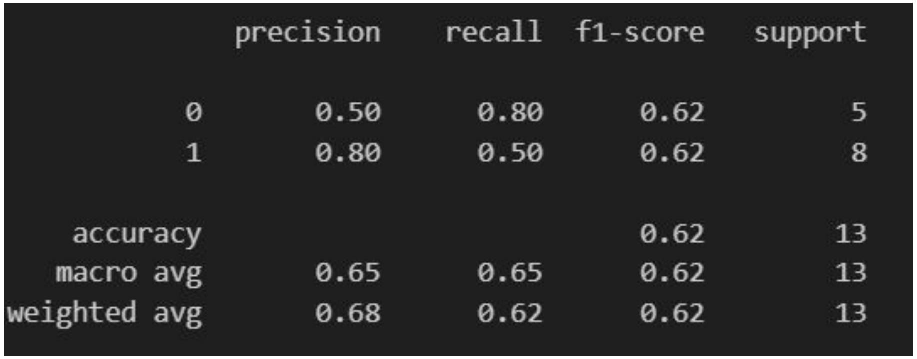


```
              precision    recall  f1-score   support

           0       0.43      0.60      0.50         5
           1       0.67      0.50      0.57         8

    accuracy                           0.54        13
   macro avg       0.55      0.55      0.54        13
weighted avg       0.58      0.54      0.54        13
```

AB

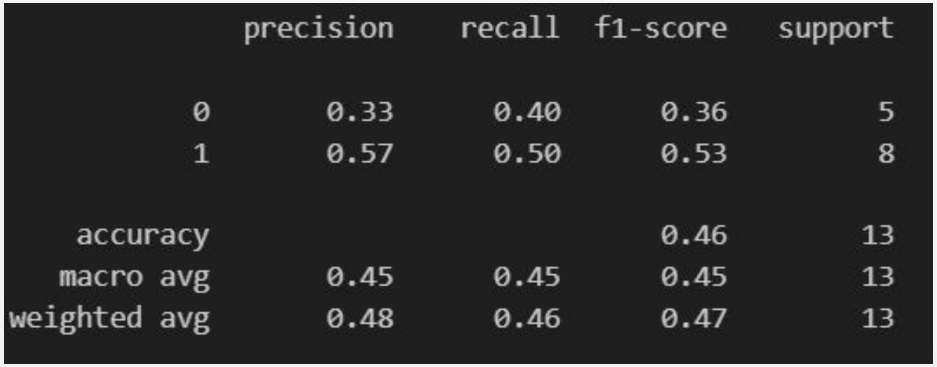


```
              precision    recall  f1-score   support

           0       0.50      0.80      0.62         5
           1       0.80      0.50      0.62         8

    accuracy                           0.62        13
   macro avg       0.65      0.65      0.62        13
weighted avg       0.68      0.62      0.62        13
```

SVM

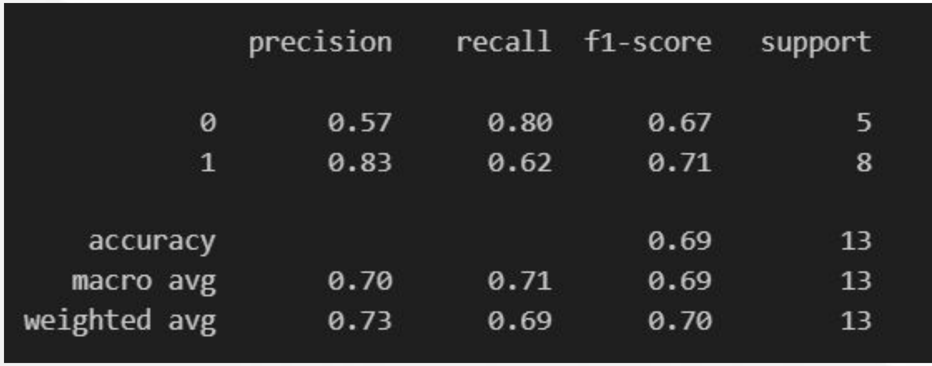


```
              precision    recall  f1-score   support

           0       0.33      0.40      0.36         5
           1       0.57      0.50      0.53         8

    accuracy                           0.46        13
   macro avg       0.45      0.45      0.45        13
weighted avg       0.48      0.46      0.47        13
```

KNN



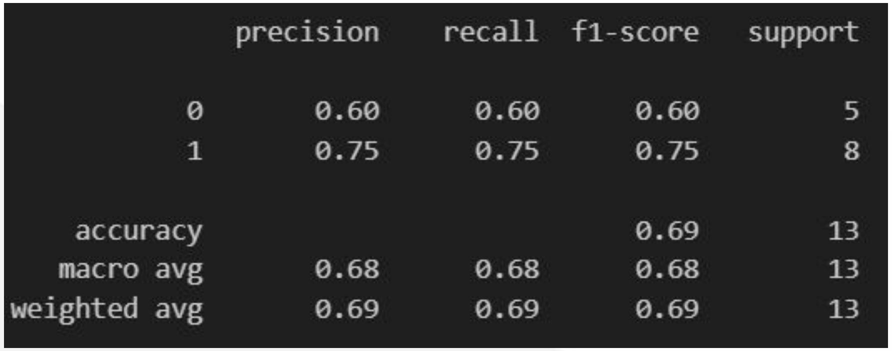
```
              precision    recall  f1-score   support

           0       0.57      0.80      0.67         5
           1       0.83      0.62      0.71         8

    accuracy                           0.69        13
   macro avg       0.70      0.71      0.69        13
weighted avg       0.73      0.69      0.70        13
```

YEAR 7

DT

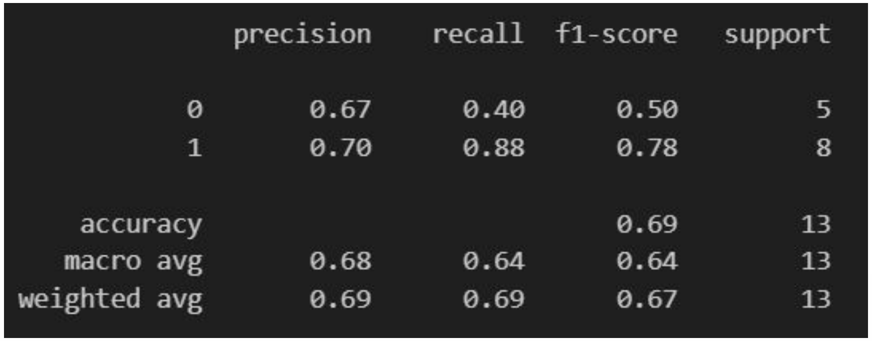


```
              precision    recall  f1-score   support

           0       0.60      0.60      0.60         5
           1       0.75      0.75      0.75         8

    accuracy                           0.69        13
   macro avg       0.68      0.68      0.68        13
weighted avg       0.69      0.69      0.69        13
```

AB

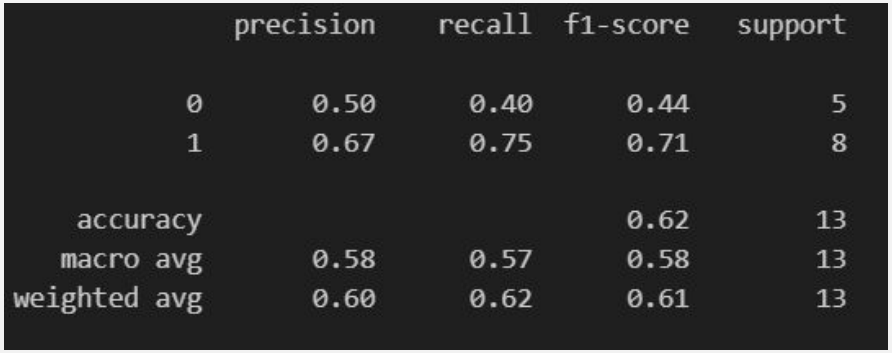


```
              precision    recall  f1-score   support

           0       0.67      0.40      0.50         5
           1       0.70      0.88      0.78         8

    accuracy                           0.69        13
   macro avg       0.68      0.64      0.64        13
weighted avg       0.69      0.69      0.67        13
```

SVM

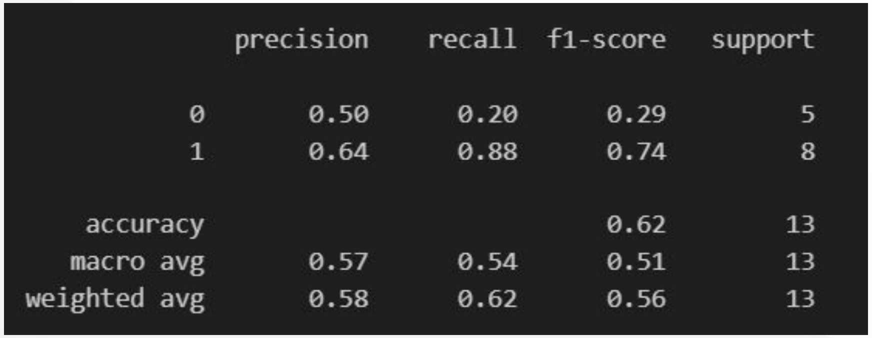


```
              precision    recall  f1-score   support

           0       0.50      0.40      0.44         5
           1       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
```

KNN



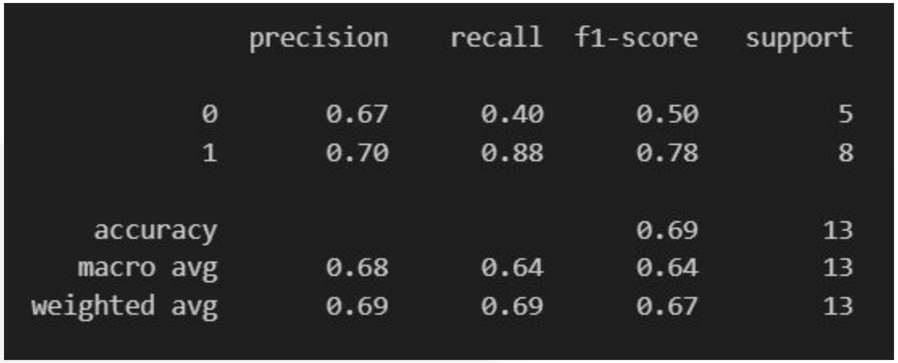
```
              precision    recall  f1-score   support

           0       0.50      0.20      0.29         5
           1       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
```

YEAR 8

DT

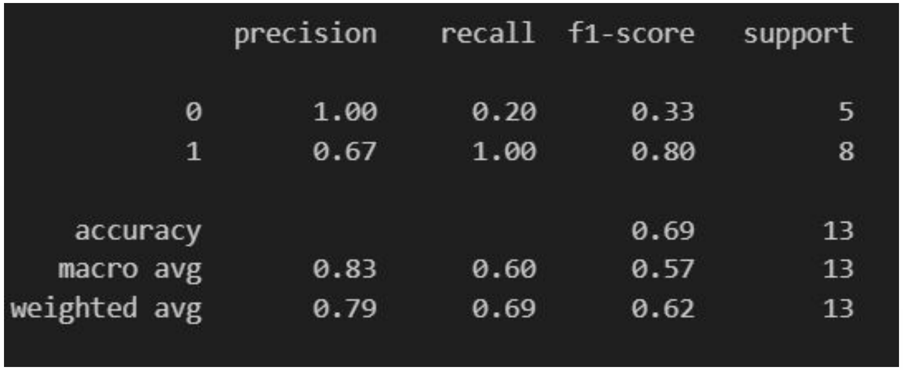


```
              precision    recall  f1-score   support

           0       0.67      0.40      0.50         5
           1       0.70      0.88      0.78         8

    accuracy                           0.69        13
   macro avg       0.68      0.64      0.64        13
weighted avg       0.69      0.69      0.67        13
```

AB

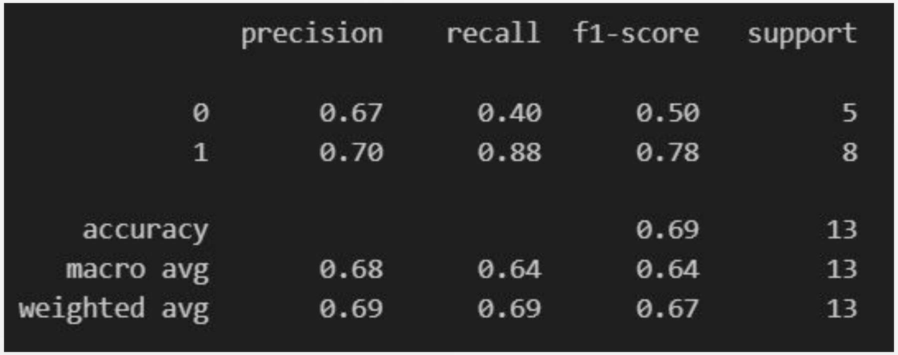


```
              precision    recall  f1-score   support

           0       1.00      0.20      0.33         5
           1       0.67      1.00      0.80         8

    accuracy                           0.69        13
   macro avg       0.83      0.60      0.57        13
weighted avg       0.79      0.69      0.62        13
```

SVM

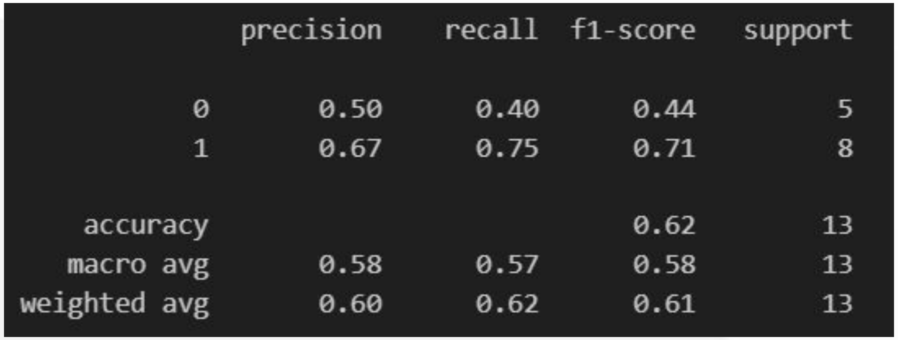


```
              precision    recall  f1-score   support

           0       0.67      0.40      0.50         5
           1       0.70      0.88      0.78         8

    accuracy                           0.69        13
   macro avg       0.68      0.64      0.64        13
weighted avg       0.69      0.69      0.67        13
```

KNN



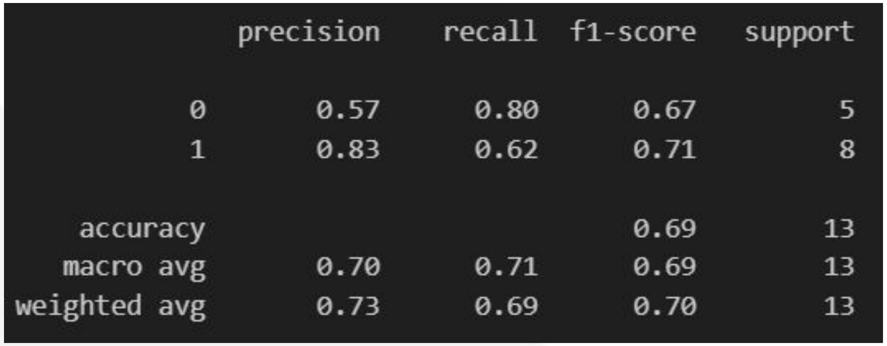
```
              precision    recall  f1-score   support

           0       0.50      0.40      0.44         5
           1       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
```

YEAR 9

DT

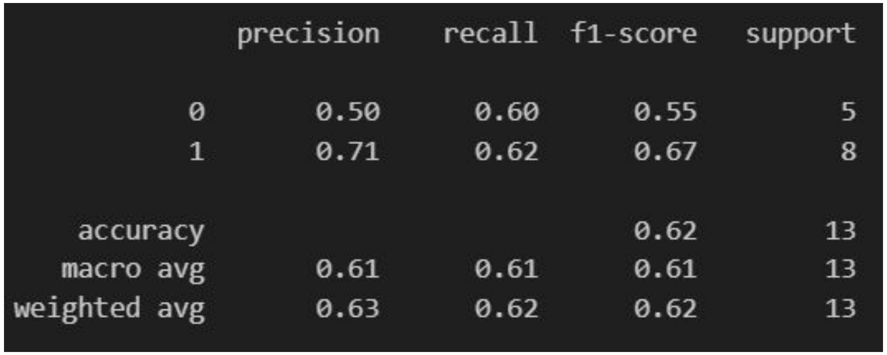


```
              precision    recall  f1-score   support

           0       0.57      0.80      0.67         5
           1       0.83      0.62      0.71         8

    accuracy                           0.69        13
   macro avg       0.70      0.71      0.69        13
weighted avg       0.73      0.69      0.70        13
```

AB

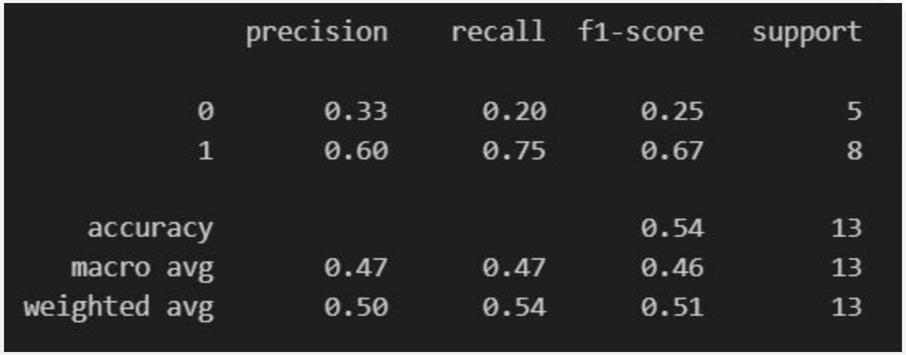


```
              precision    recall  f1-score   support

           0       0.50      0.60      0.55         5
           1       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
```

SVM

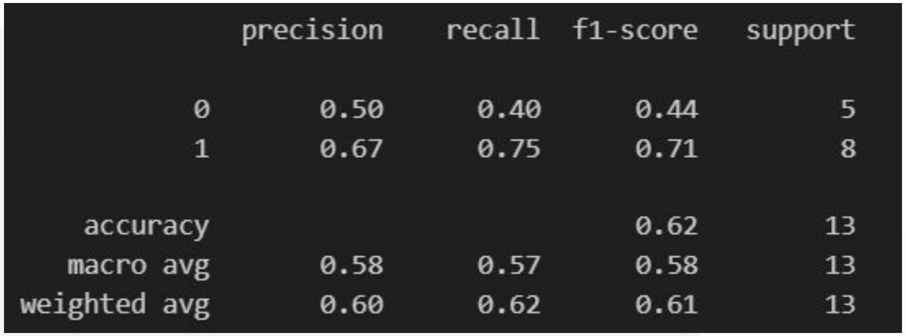


```
              precision    recall  f1-score   support

           0       0.33      0.20      0.25         5
           1       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
```

KNN



```
              precision    recall  f1-score   support

           0       0.50      0.40      0.44         5
           1       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
```