

Aprendizagem Computacional (M.EIC001)
2023/2024

Basketball Playoffs

a Data Mining Project

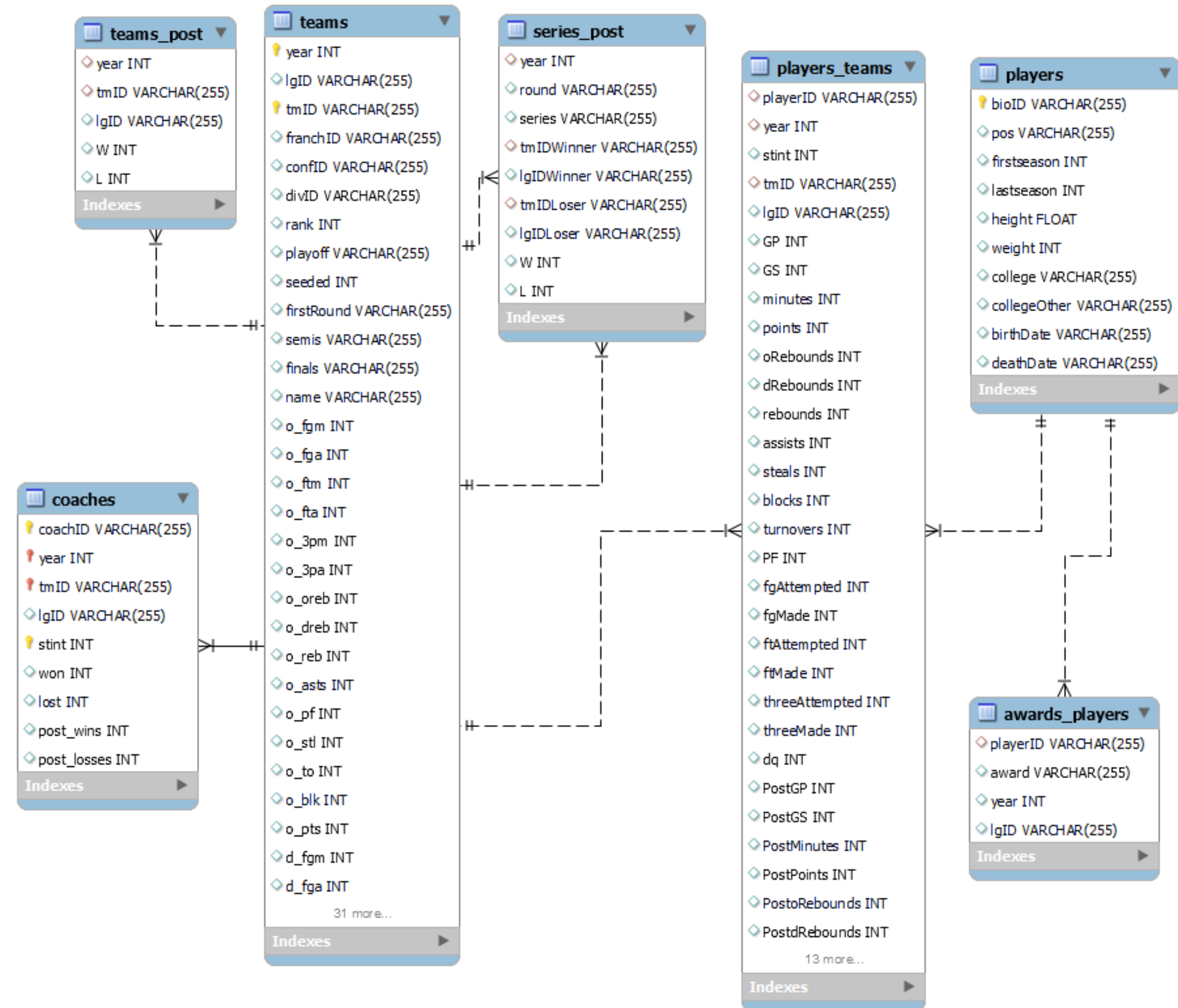
G72

António Ferreira - up202004735
João Maldonado - up202004244
Tomás Gomes - up202004393

Domain Description

For this project, we're working with a 10 year span of WNBA data. Specifically:

- Players and their annual statistics (including playoffs' statistics)
- Coaches
- Teams and their yearly performance (including playoffs' performance)
- Awards



(fig. 1) Entity Relationship Diagram

Exploratory Data Analysis

For this part, we focused on analyzing the given dataset, in order to better understand the data. During this, we analyzed:

Target Variable Distribution

We came to the conclusion that the dataset is **balanced** regarding the target variable - **playoffs**.

- In total, we verified that, in 142 entries, **80 teams passed to playoffs**, while **62 didn't**.

Inconsistent Data

We tried to find data that was violating the **patterns** and **rules** of the provided context.

- For instance, we found that, in the sixth year, both **Connecticut Sun** and **Sacramento Monarchs** won the playoffs, which can not happen.

Exploratory Data Analysis

For this part, we focused on analyzing the given dataset, in order to better understand the data. During this, we analyzed:

Missing Values

Our main focus was trying to find any data record that had missing values.

- In the **players table**, there's many entries that had **all of their values missing**. We later discovered that **52** of these records corresponded to coaches.
- Besides that, we also found that, in the players table, many rows had their weights and heights missing.

Outliers


In this case, we searched for values that were outside of the normal range of a given feature.

- There's a player with a **height of 9.0**, which is way below the average height values of the other players.

Predictive Data Mining Problem

Problem Definition

Determining which **8 teams** (4 from the East Conference and 4 from the West Conference) qualify to the playoffs, based only on data from the previous 10 years.



Predictive Data Mining Problem

Data Preparation

Missing Values

- Fixed the missing weight and height values, of the players (not coaches). For this we used **Mean Imputation**, grouping the players by position.

Outliers

- Replaced the height of the player with the average height of the players, that play in her position.

Data Inconsistencies

- Fixed an award's name
 - Kim Perrot Sportmanship → Kim Perrot Sportmanship Award
- Considered only how far a team reached, instead of the taking into account if they won a specific round.

Predictive Data Mining Problem

Data Preparation

Attribute Creation

- **Weighted average performance** of the teams, throughout the years. In this case, a team's performance is the sum of the statistics of the players that play in it.
- **Average weight and height** of the teams
- **Number of awards** the players and the coach obtained
- **Number of times** the players of a team reached the **playoffs, semi-finals and finals**
- **Coach win ratio**

Attribute Selection

- We decided to include:
 - **year**
 - **conference ID**
 - **coach stint**
 - **playoff**

Predictive Data Mining Problem

Experimental Setup

For our model, we divided the data into train and test sets with the following logic:

- in the train sets, we included data from the years 2 through 9.
- in the test sets, we included the year 10 data.

Note

Evidently, in the first year, there's no historical data, for the model to be trained on. Therefore, we don't train our predictive model on this years' data.

Predictive Data Mining Problem

Results

For this problem we tried many algorithms in order to obtain the best possible results. Between all of them, the algorithm that showed the best results was: **Logistic Regression**.

Statistics

- **Accuracy - 84.6%**
- **Precision - 87.5%**
- **Recall - 87.5%**
- **F1-Score - 87.5%**

Predictions

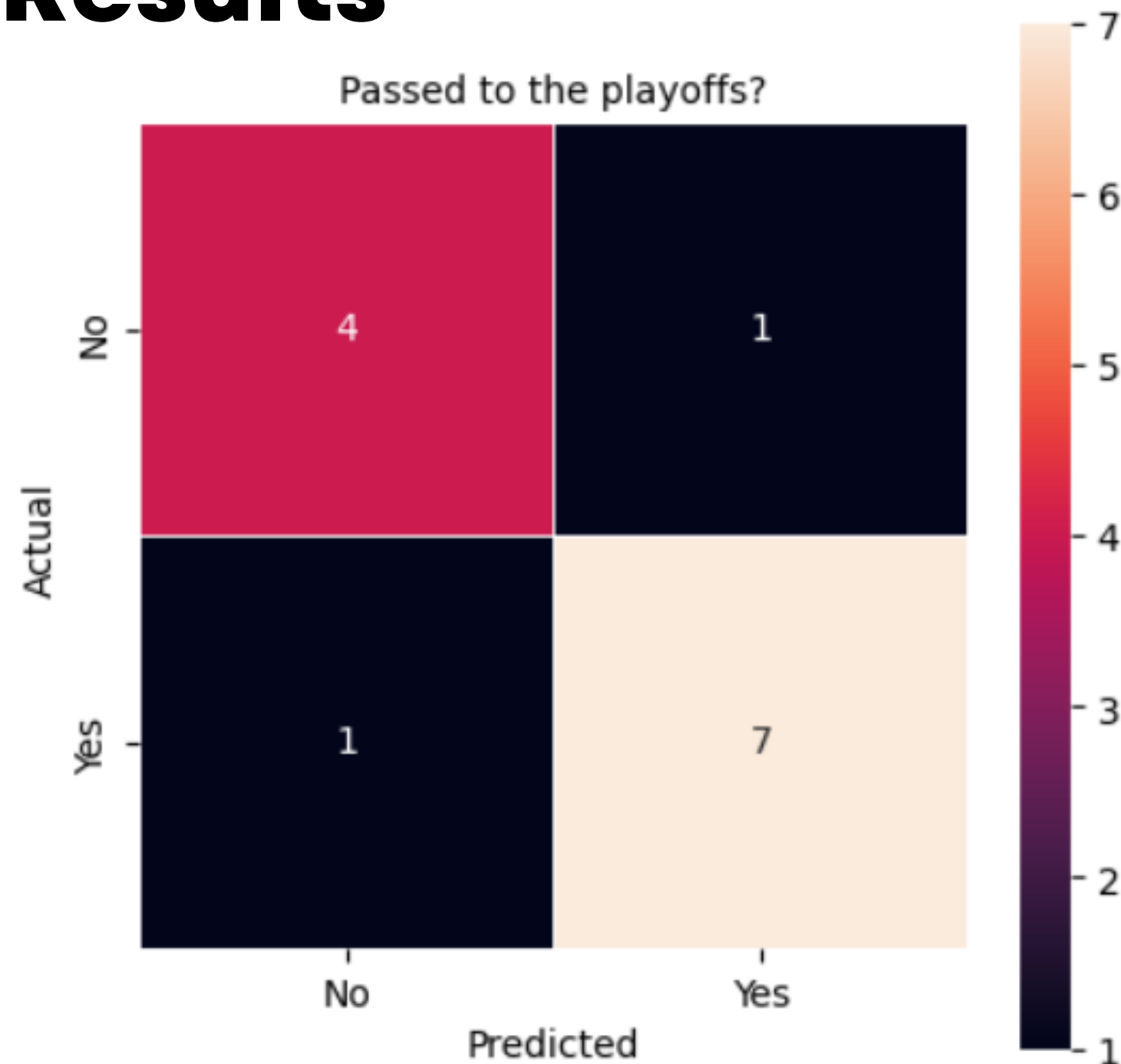
The values, regarding the statistics shown, were obtained without taking into consideration the problem's context. Therefore, it's our job to handle which 4 teams, from each conference qualify for the playoffs.

West Conference - LAS, PHO, SAS, SEA

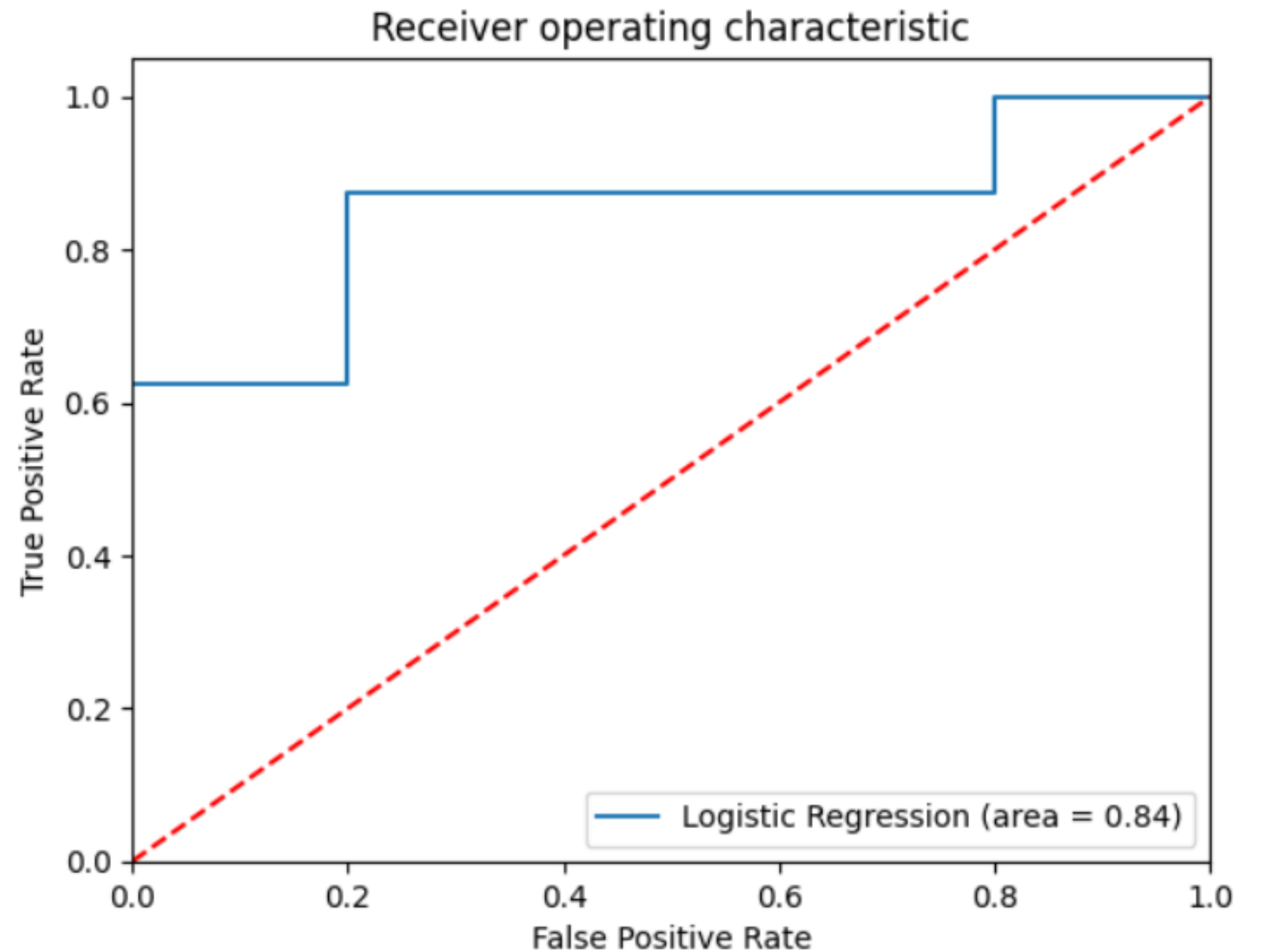
East Conference - DET, IND, CHI, ATL

Predictive Data Mining Problem

Results



(fig. 2) Confusion Matrix



(fig. 3) ROC curve and AUC score

Conclusions

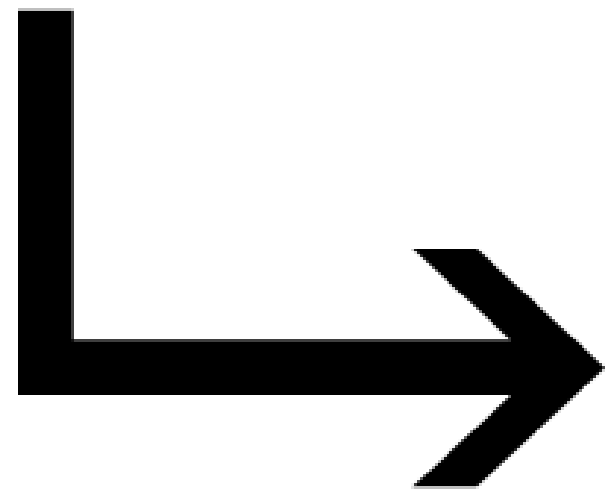
Considering the problem's context, in the sports world, it's impossible to always accurately predict the outcome of a season.

On the other hand, there are many ways to tackle this problem, which means there is no approach that is correct and efficient for every scenario.

In the future, to better our model, we can **create new features**, like the **number of times each coach reached the playoffs, the semi-finals and the finals**, or give **different weights to each award**, according to their importance.

Attachments

In the following set of slides, we'll be diving deeper into the many processes and steps that weren't mentioned during the main presentation, due to the limitation of slides and time constraints.



More specifically, our objective is not only to extensively **document our progress** through out the weeks, but also **present the other models** we created, and their **respective results**.

A. Business Understanding

Analysis of requirements with the end user

After analyzing the given data, we've come to the conclusion that the user should be able to predict, with the best precision possible, which 8 teams qualify to the playoffs. This decision is based on data available in the beginning of the current season which includes the teams composition (players and coach), players physical characteristics and historical data regarding their performance.

Definition of business goals

When considering the business aspect of this project, we can consider two main goals:

- **Qualification Prediction** - The primary goal of this project is to accurately predict which teams make the playoffs in the next season. This can be valuable for team management, betting agencies, sports analysts, and fans.
- **Team Strategy Optimization** - Many patterns and trends can be discovered by studying the performance of each team over the course of ten years. Using this data, each team can improve its player selection and performance, increasing its chances of making the playoffs.

A. Business Understanding

Translation of business goals into data mining goals

Based on the business goals we have specified earlier, we can easily identify many data mining goals for this project:

- **Historical Data Analysis** - By analyzing the data of the given 10 years, we can identify the key factors that contributed for a team's qualification.
- **Predictive Modeling** - With the data, we can create a model which will determine which teams qualify for the playoffs.
- **Pattern Recognition** - We can identify patterns in player performances and other statistics.
- **Feature Engineering** - Determining which variables are most predictive of playoff qualification.
- **Model Validation and Refinement** - Testing and refining the predictive model on a continuous basis to ensure accuracy and reliability over time.

A. Business Understanding

With this in mind, during the first week of the project, our main task was to **understand the problem's objective and the various concepts that revolved around it**. More specifically, while we had a general understanding of basketball, there were many terms in the given dataset that we were unfamiliar with.

Therefore, we created a text file with a list made up of terms and their respective definition.

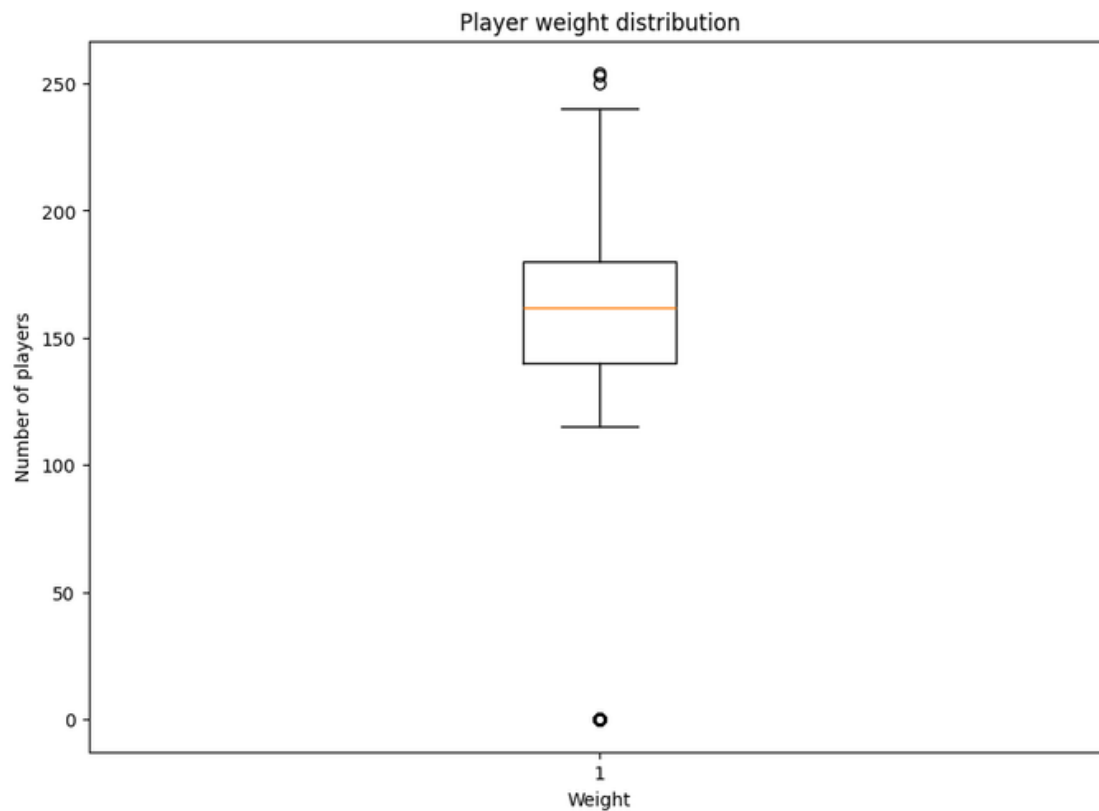
B. Pipeline Creation

At this stage, we began to analyze and study the provided data in order to identify potential hidden relationships and outliers. To make our job easier, we did the following:

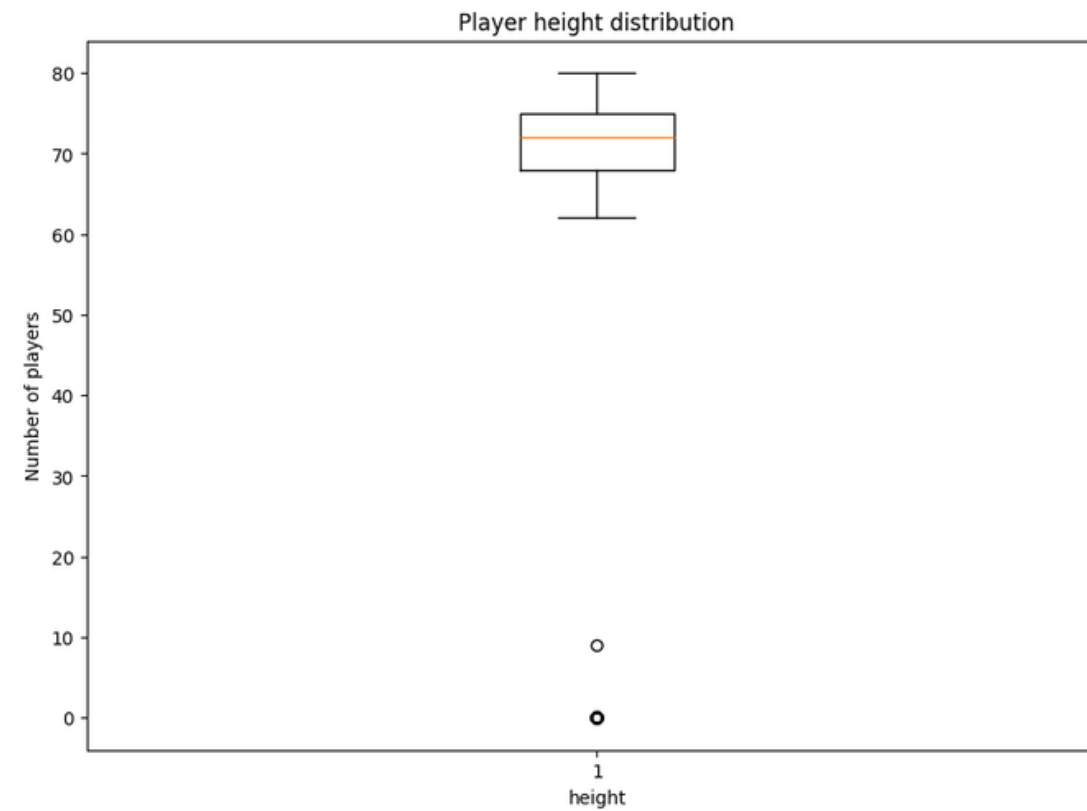
- We converted the **.csv files** into a **Relational Database**. We decided to use **SQLite**, due to its speed and portability.
- We identified the target variable: **playoff** (from the 'Teams' table)
- Using **SQLiteStudio**, we converted the original data files. However, the resulting tables had all their columns as a 'Text' type. As a result, we converted each column to their original respective types. We also added the various foreign and primary keys, according to the UML diagram.

Using the database we created, we connected to it which allowed us to retrieve data from the database and use it to begin building our models. We used the SQLite3 Python library, which connected directly to our SQL database, for this. To test it, we selected columns from the database, split the data into training and test sets, and ran a Decision Tree Classifier algorithm.

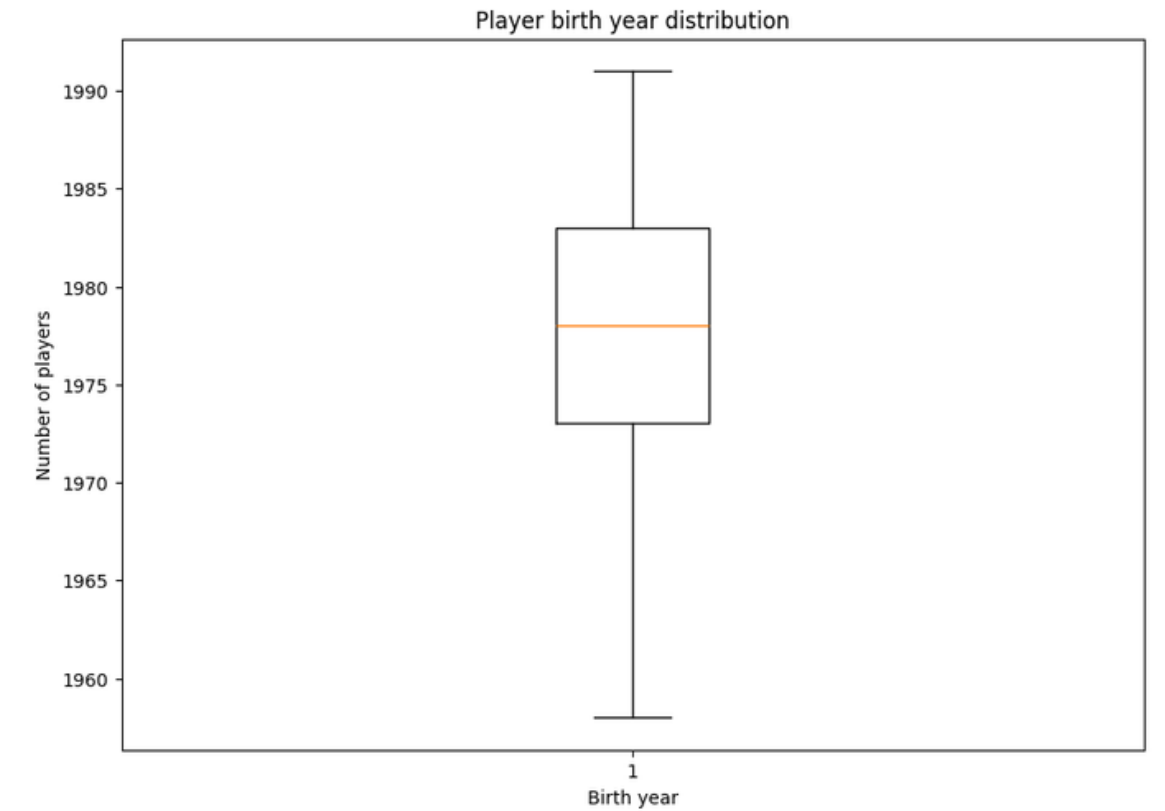
C. Data Understanding



(fig. 4) Player weight distribution



(fig. 5) Player height distribution

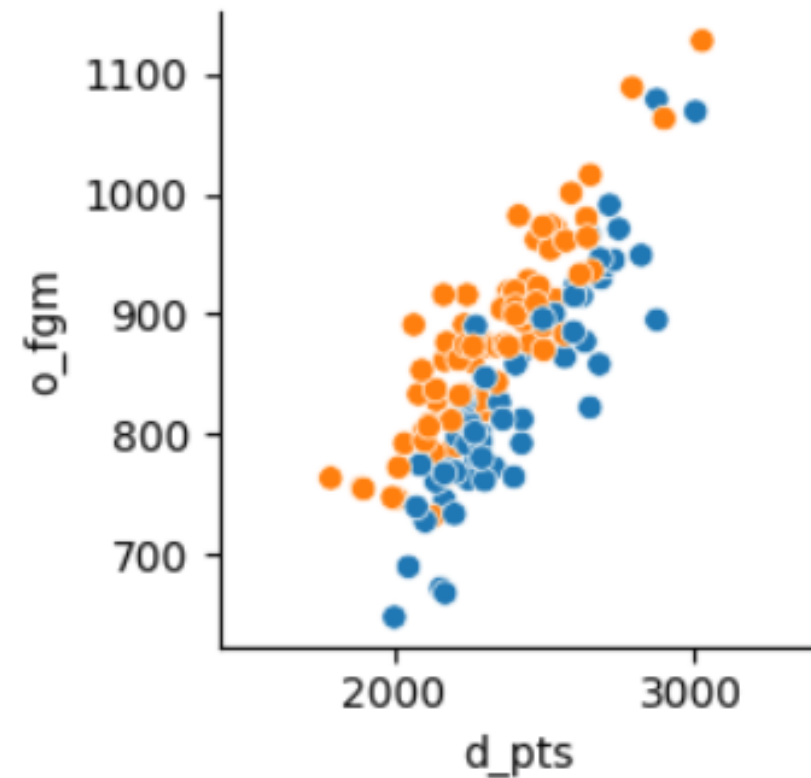


(fig. 6) Player birth year distribution

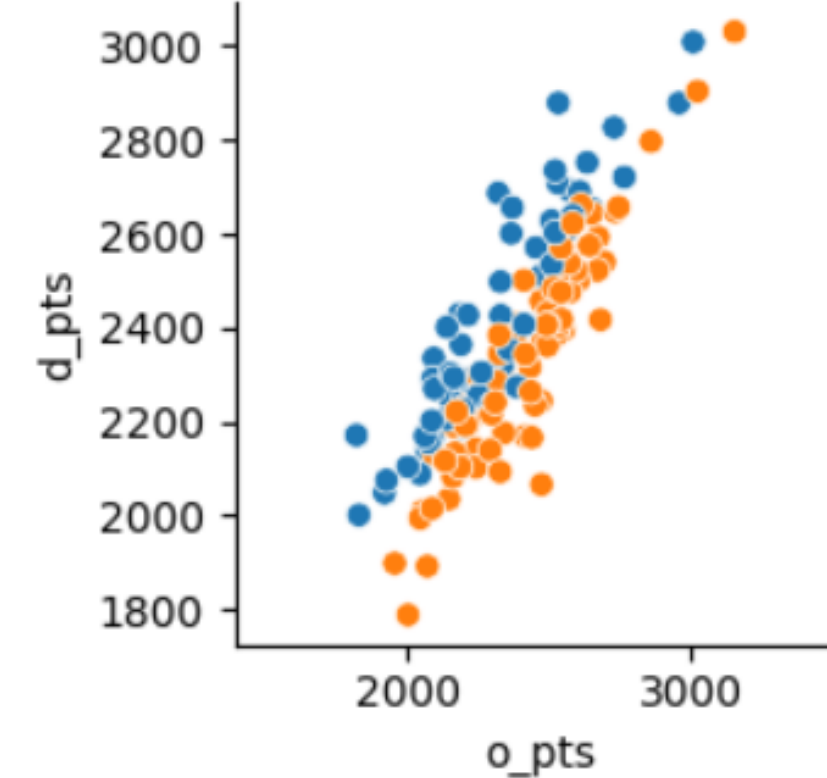
In the context of our basketball dataset, it is crucial to examine the variables of **height, weight, and birth year** for the players. Upon plotting the distribution of these variables, certain observations can be made:

- The weight distribution reveals **a mean of 150 lbs**, with notable outliers. Anomalies are observed at 0 lbs, which are clearly incorrect, whereas outliers at 250 lbs are accurate.
- The height distribution centers around a **mean of 70 inches (177 cm)**, a typical range for a feminine dataset. However, there are outliers at 0 and 10 inches, indicating erroneous data entries.
- As for the players' ages, the mean is 43 years, suggesting that **the dataset may not be recent**. Typically, basketball players fall within the age range of 20 to 37 years.

C. Data Understanding



(fig. 7) Relation between o_fgm and d_pts



(fig. 7) Relation between o_pts and d_pts

In order to understand this dataset patterns, we analysed the relation between the several columns of teams. After analysing the graph, there were two relations that kept our attention - “o_fgm” between “d_pts” and “o_fgm” between “d_pts”.

As seen in figure 7, this pattern suggests a trend where teams with higher field goals made relative to their defensive points tend to qualify for the playoffs.

In figure 8, the pattern suggests a trend where teams with higher offensive points relative to their defensive points tend to qualify for the playoffs.

D. Data Processing

In this step, besides the mentioned steps in the presentation, we also did the following:

Data conversion for compability with MLP algorithms

Our data set is made up of many categorical fields. As a consequence, one of our main tasks in this stage of the project was dealing with this type of data, in order to use these fields in our models.

To combat this, we created dummy variables (One-Hot Encoding). In this process, for each unique value of a categorical value, a binary value (0 or 1) is created. If a value was found in a row in the original column, the corresponding new column will have a 1, otherwise it will have a 0.

Data scaling

This dataset is composed by a great number of features that appear in multiple scales. Therefore, it is important to scale all this features to a common range. With this in mind, we applied a Min Max Scaler to our input data, so that every feature value ranges between 0 and 1.

D. Data Processing

Feature Engineering

In our most **complex models**, we created various new features with different complexities. The **simpler features** include:

- *the sum of the number of times the players of a team reached the playoffs, semis and finals;*
- *the sum of the number of awards the players and coaches of a team won;*
- *the coach win ratio.*

There are also more **complex features** like:

- *the weighted average of the players statistics, like points and rebounds, where the most recent years have a greater weight;*
- *when the players are rookies the previously mentioned weighted average is replaced with the average of the rookies from the previous year that play in the same position.*

E. Data Modelling

At this stage, our main focus was on **testing various models**, in order to achieve the best results possible. Knowing this, we created **four distinct models**, each with varying degrees of complexity:

1. **“The most basic model possible”**
2. **“The team performance model”**
3. **“The model based on the players’ performance prediction”**
4. **“The best model”**

E. Data Modelling

“The most basic model possible”

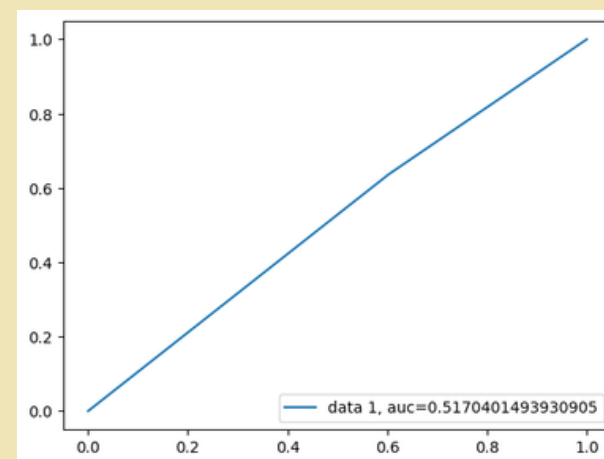
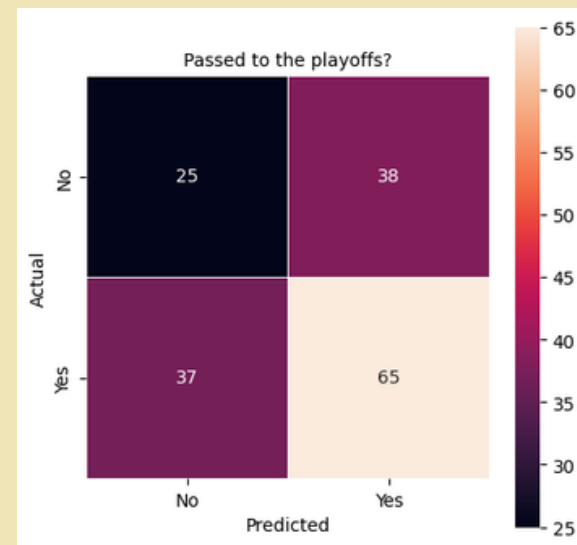
This was the first model we created, and it served as a **foundation for testing various machine learning algorithms**. We only chose the **team ID, player ID, year, and playoff** (whether the team made the playoffs or not) for this model.

The data was then divided into training (**the first nine years of data**) and test sets (**the tenth year of data**). Also, due to the inclusion of the player IDs, we aren't determining which teams go to the playoffs, but which players go instead.

E. Data Modelling

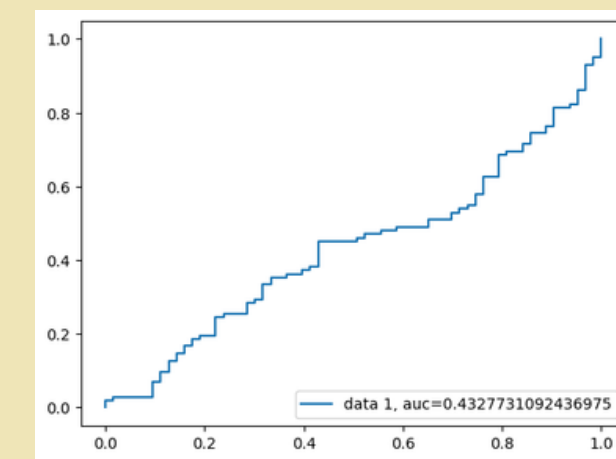
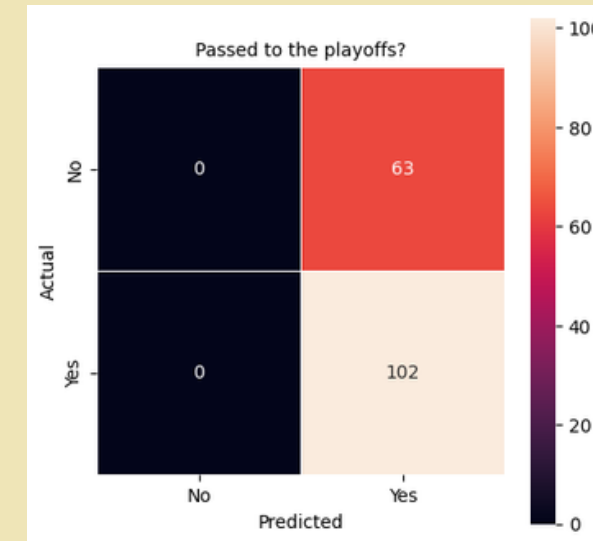
Decision Tree Classifier

Accuracy: 0.545
Precision: 0.631
Recall: 0.637
F1: 0.634



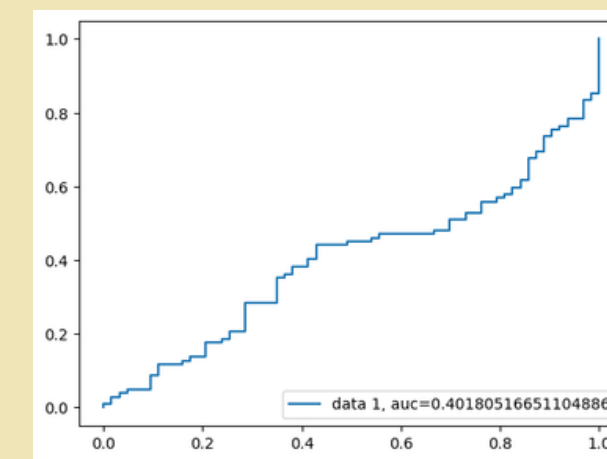
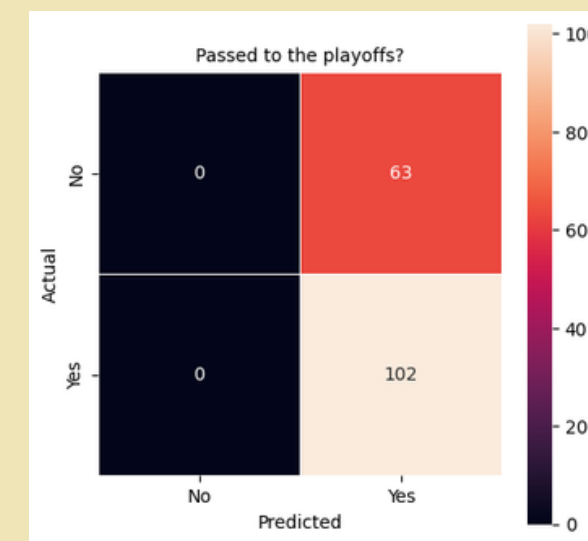
Logistic Regression

Accuracy: 0.62
Precision: 0.62
Recall: 1.00
F1: 0.76



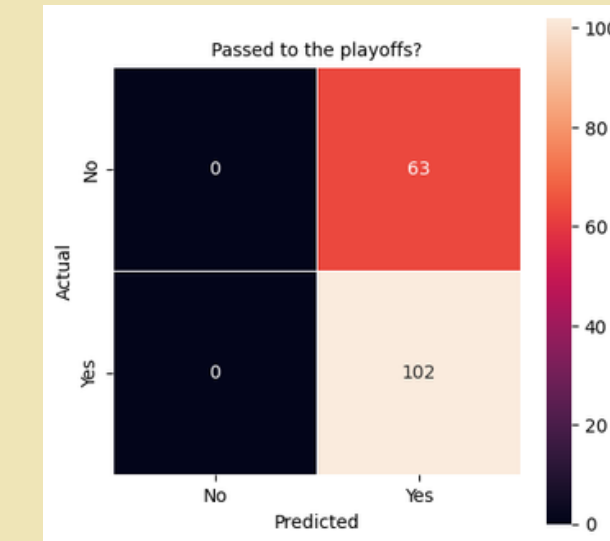
Naïve Bayes

Accuracy: 0.62
Precision: 0.62
Recall: 1.00
F1: 0.76



Support Vector Machine (SVM)

Accuracy: 0.62
Precision: 0.62
Recall: 1.00
F1: 0.76



E. Data Modelling

The results of the four algorithms chosen show that, on one hand, **the obtained values aren't great**, which is explained by the small number of attributes chosen.

These algorithms, on the other hand, don't tell us anything, because they **don't determine which teams make the playoffs**. They were only used as a testing base.

E. Data Modelling

“The team performance model”

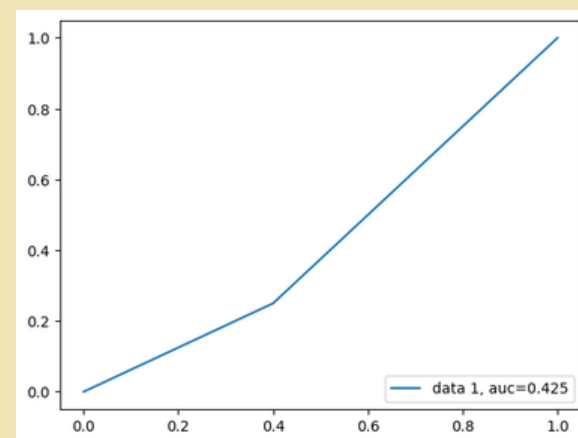
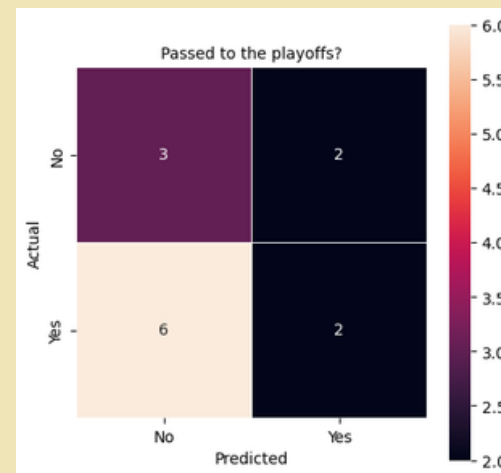
We attempted to focus on a key aspect of this project in this second model: **a team is made up of players**. As a result, in addition to the team ID, year, and playoff attributes, we included the average performance of each teams' players, in each year. We also added the yearly average heights and weights of each team, the win and loss ratio, the number of player awards, the number of coach awards and if a team reached the semifinals and the finals.

It also should be noted that, once again, the train-test split done in this model was the same as the previous one.

E. Data Modelling

Decision Tree Classifier

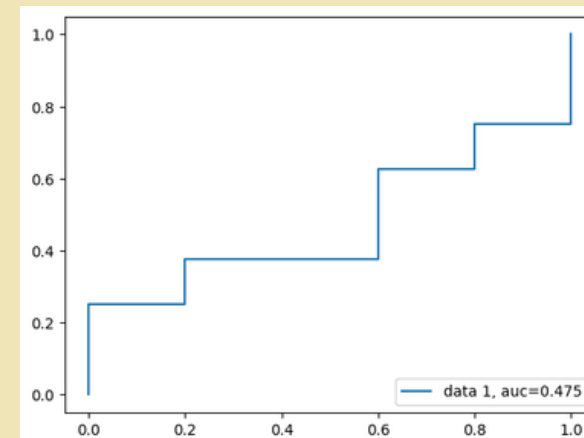
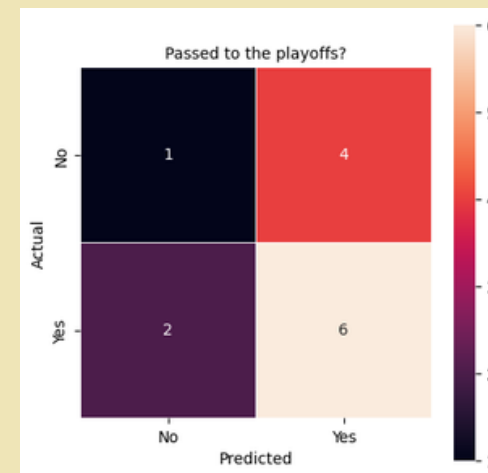
Accuracy: 0.384
Precision: 0.5
Recall: 0.25
F1: 0.333



Qualified Teams:
LAS, MIN, PHO, SAS, CON, NYL, WAS, ATL

Logistic Regression

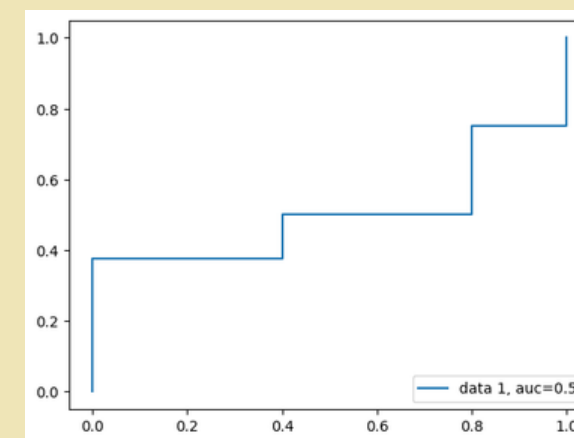
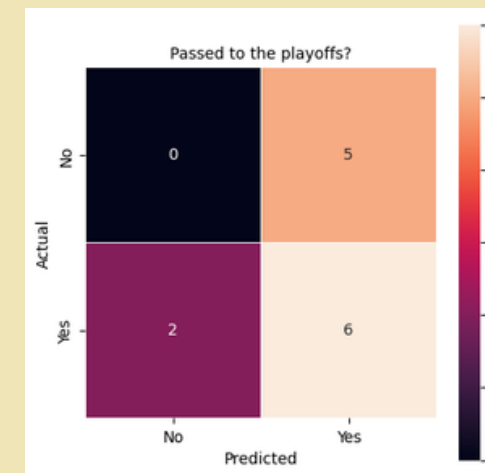
Accuracy: 0.54
Precision: 0.60
Recall: 0.75
F1: 0.67



Qualified Teams:
LAS, SEA, SAC, SAS, DET, CON, NYL, WAS

Naive Bayes

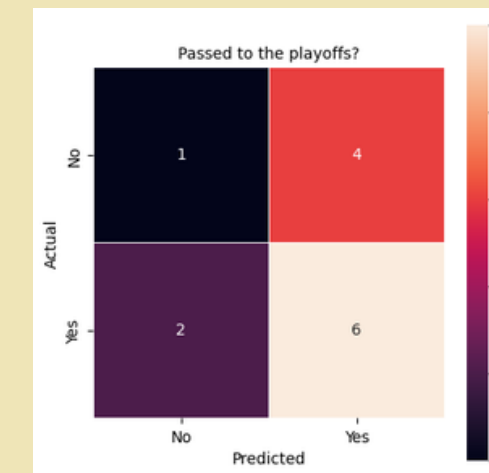
Accuracy: 0.46
Precision: 0.55
Recall: 0.75
F1: 0.63



Qualified Teams:
SAS, LAS, SAC, SEA, DET, NYL, CON, IND

Support Vector Machine (SVM)

Accuracy: 0.54
Precision: 0.60
Recall: 0.75
F1: 0.67



Qualified Teams:
LAS, SEA, SAS, PHO, DET, CON, NYL, CHI

E. Data Modelling

The most important takeaway from the model's results is that we can finally **predict which teams will qualify**. However, it is important to note that the model does not account for the following constraint: **only eight teams can make the playoffs, with four teams from the East conference and the other four from the West conference**. This can be seen in the results (for example, in the confusion matrices).

As it is stated in the presentation, the process to determine if a team passed, was centered around the **probabilities calculated during the execution of each algorithm**. Even if a team's probability was below 50%, if it was in the **top 4 of its conference**, then it made the playoffs.

E. Data Modelling

“The model based on the players’ performance prediction”

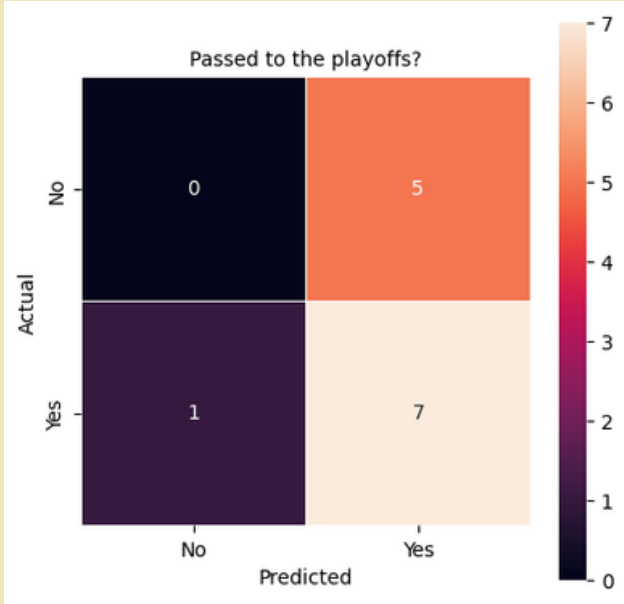
Although the previous model produced some promising results, its poor execution could be due to the treatment of each team as an **entity rather than a group of players**. To counteract this, we shifted our focus to **each player's performance**.

With this in mind, the main idea behind this new model was to **predict each player's performance**, in a year, based on previous years' performance and then **rank the teams based on their overall predicted performances**. The **four teams** from each conference with the highest score qualify.

E. Data Modelling

Logistic Regression

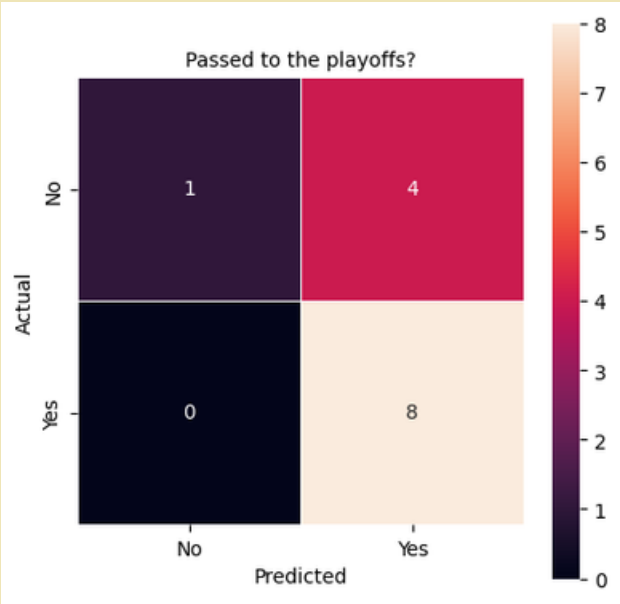
Accuracy: 0.5384615384615384
Precision: 0.5833333333333334
Recall: 0.875
F1: 0.7000000000000001



Qualified Teams:
SAC, SEA, SAS, LAS, CON, NYL, WAS, CHI

Neural Network

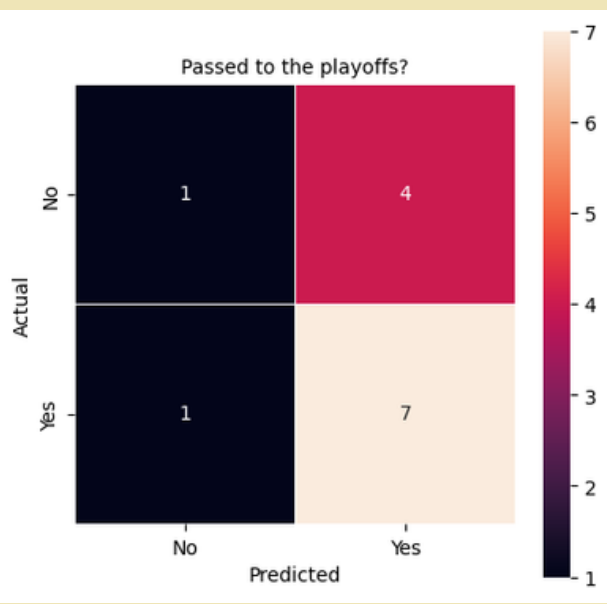
Accuracy: 0.6923076923076923
Precision: 0.6666666666666666
Recall: 1.0
F1: 0.8



Qualified Teams:
SAS, SEA, LAS, SAC, ATL, CHI, WAS, CON

Support Vector Machine (SVM)

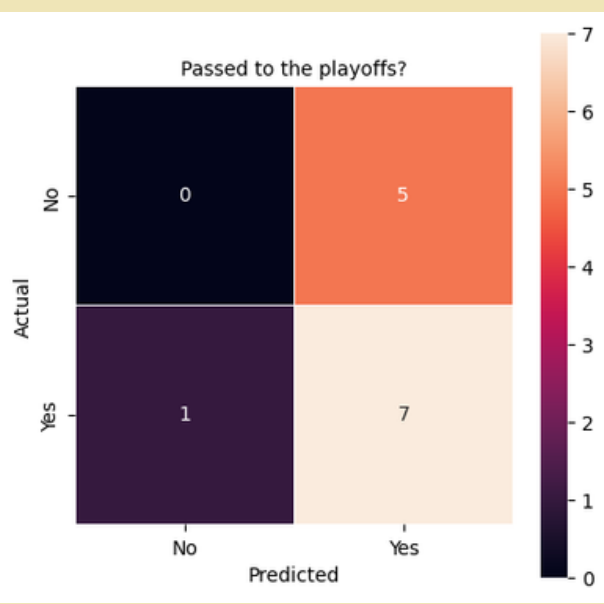
Accuracy: 0.6153846153846154
Precision: 0.6363636363636364
Recall: 0.875
F1: 0.7368421052631579



Qualified Teams:
SAC, SEA, SAS, LAS, CON, NYL, WAS, CHI

K-Nearest Neighbour

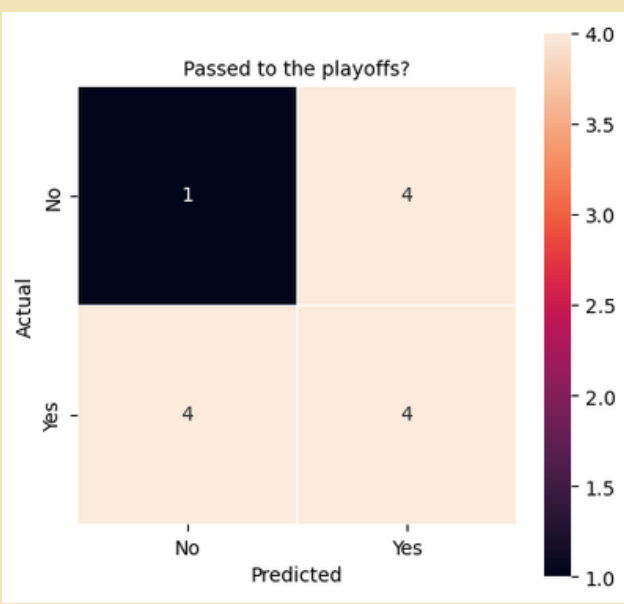
Accuracy: 0.5384615384615384
Precision: 0.5833333333333334
Recall: 0.875
F1: 0.7368421052631579



Qualified Teams:
SEA, SAC, SAS, MIN, CON, WAS, DET, NYL

Decision Tree Classifier

Accuracy: 0.38461538461538464
Precision: 0.5
Recall: 0.5
F1: 0.5



E. Data Modelling

From the graphics shown, it's easy to see that there wasn't much of an **increase** in the quality of the results, as the **accuracy** and **precision** values are **similar (or lower)**, when compared to previous models.

This is most likely due to the fact that we are **predicting each player's performance**. This task not only is **inherently complex**, leading to poor results, but also it's important to note that a performance of a player, from one year to another, can be **affected by various external factors**, which we can't take into consideration.

E. Data Modelling

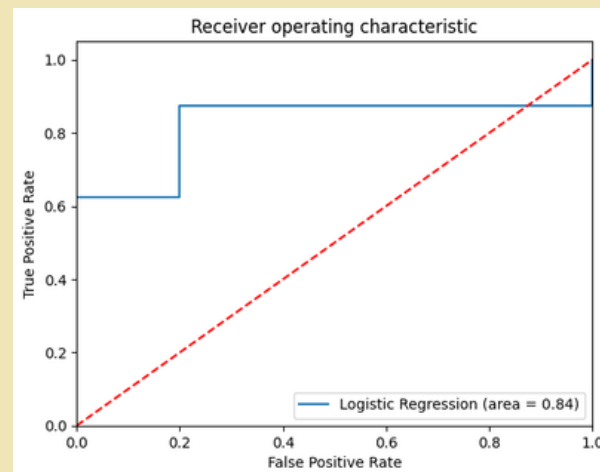
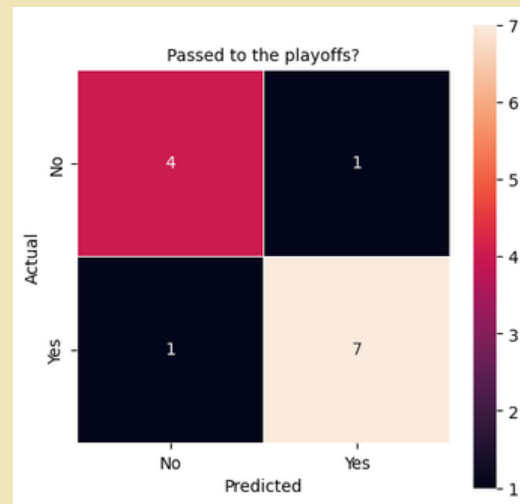
“The best model”

This was the model that was showcased in the presentation. Even though it was already described, in this section we'll include further information regarding it, like the results of the other algorithms tested.

E. Data Modelling

Logistic Regression

Accuracy: 0.8461538461538461
Precision: 0.875
Recall: 0.875
F1: 0.875

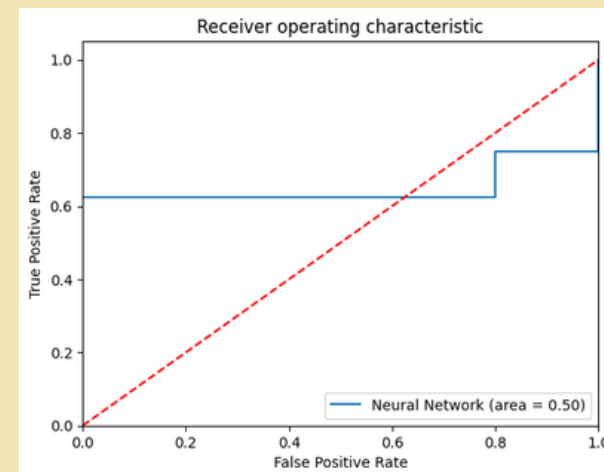
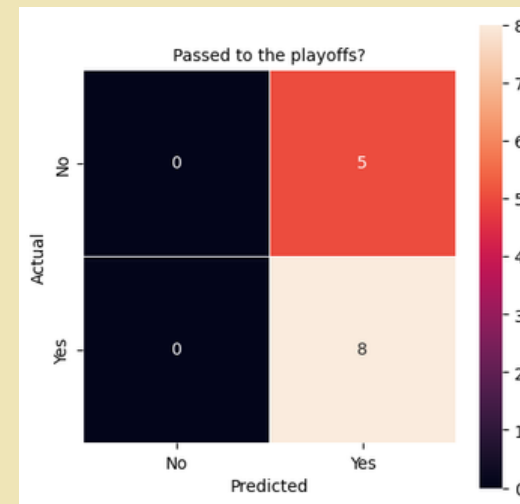


Qualified Teams:

LAS, PHO, SAS, SEA, IND, DET, CHI, ATL

Neural Network

Accuracy: 0.6153846153846154
Precision: 0.6153846153846154
Recall: 1.0
F1: 0.761904761904762

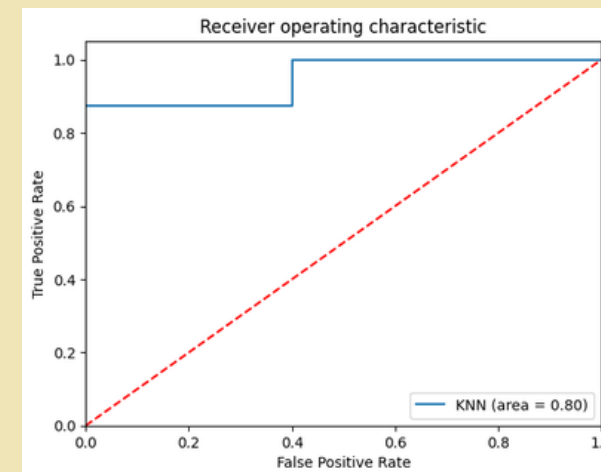
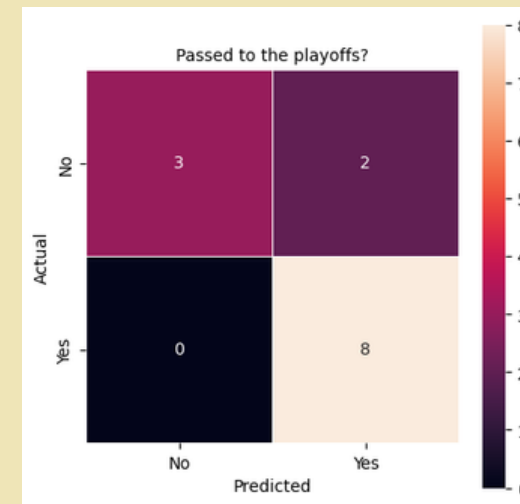


Qualified Teams:

LAS, SAS, PHO, SAC, DET, CHI, IND, ATL

K-Nearest Neighbours

Accuracy: 0.8461538461538461
Precision: 0.8
Recall: 1.0
F1: 0.888888888888889

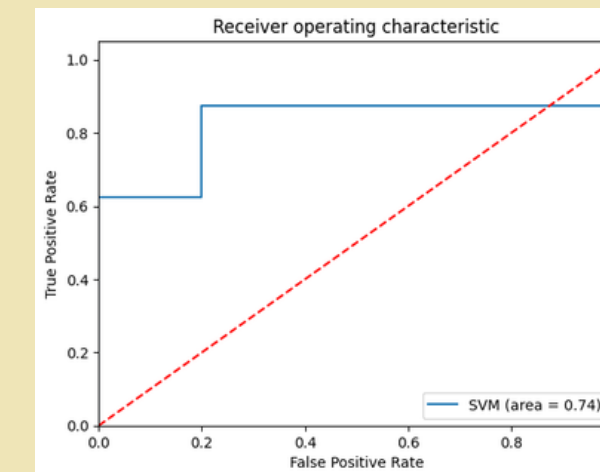
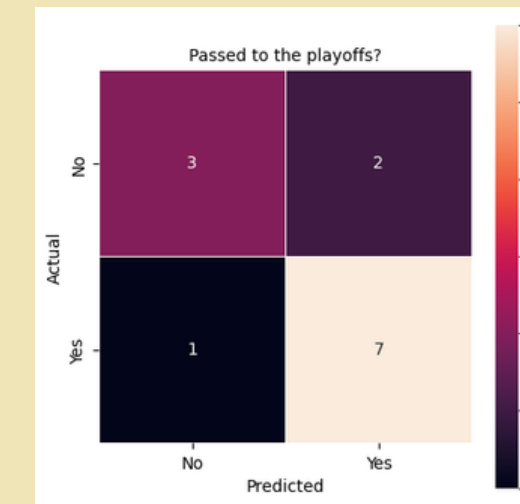


Qualified Teams:

SAS, LAS, SEA, PHO, DET, IND, ATL, CHI

Support Vector Machines (SVM)

Accuracy: 0.7692307692307693
Precision: 0.7777777777777778
Recall: 0.875
F1: 0.823529411764706



Qualified Teams:

LAS, PHO, SAS, SEA, DET, IND, CHI, ATL

E. Data Modelling

From the results shown, we can easily conclude that this is the best model, since it has the best results possible. More specifically, the **Logistic Regression** version, since it produced the best values in all the metrics. Therefore, this is the model we used for the competition.

F. Competition

After receiving the data regarding the season 11, we applied our chosen model and obtained the following results:

Team	Playoff
LAS	Y
SEA	Y
PHO	Y
SAS	Y
TUL	N
MIN	N

Team	Playoff
IND	Y
NYL	Y
CON	Y
ATL	Y
WAS	N
CHI	N