WNBA PLAYOFF Prediction

G22 - 2023/2024

Gustavo Costa - up202004187
João Oliveira - up202004407
Ricardo Cavalheiro - up202005103

Table of contents

Business Understanding

Analysis of Requirements with End User

- **Investors** & **stakeholders** are looking for the best teams to invest their funds in order to maximize their returns.
- **Teams analysts'** want to know which statistics have the most impact in the team performance and also the ones they need to improve at.

Business Understanding

Business Goals

- Successfully predict the playoff qualification of at least 70% of the teams.
- The project must be completed until its due date.

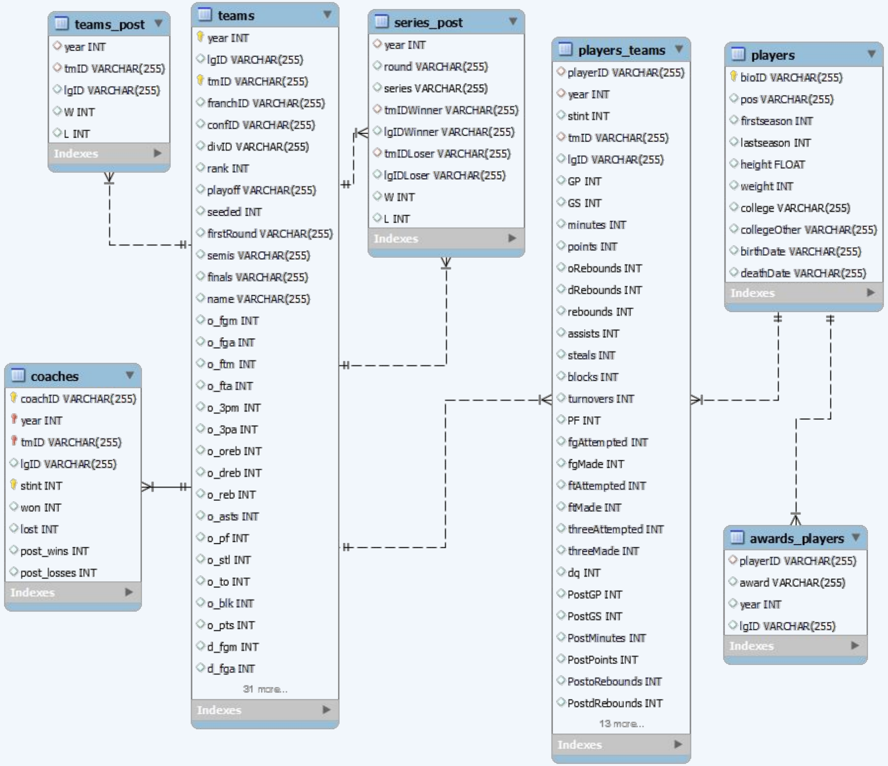
Business Goals > DM Goals

- Building a model to predict whether or not a team will qualify for the playoffs.
- Obtain an accuracy of at least 70% & AUC over 0.8.

Domain Understanding

10 Years Data of WNBA Seasons

- **Teams -** 143 entries
- **Players -** 894 entries
- **Players_Teams -** 1877 entries
- **Coaches -** 163 entries
- **Awards_Players -** 96 entries
- **Teams_Post -** 81 entries
- **Series_Post -** 71 entries



Exploratory Data Analysis

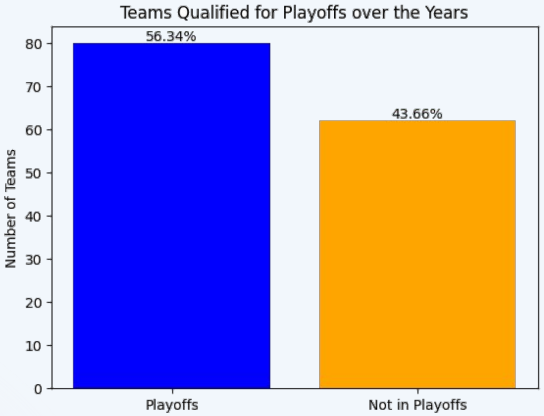


Fig 1 - Target Distribution

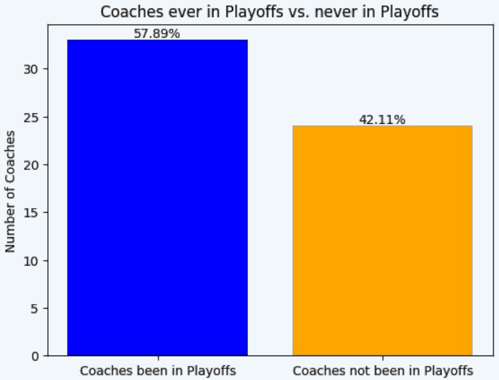


Fig 2 - Coaches Playoff Appearances Distribution

Exploratory Data Analysis

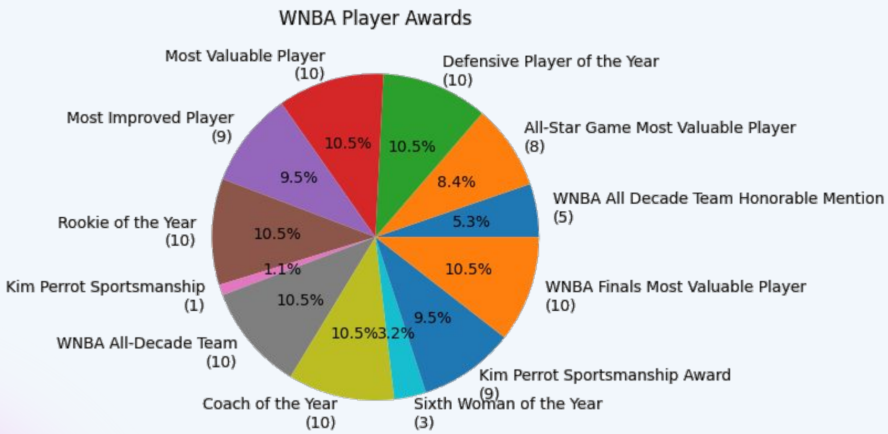


Fig 3 - Awards Distribution

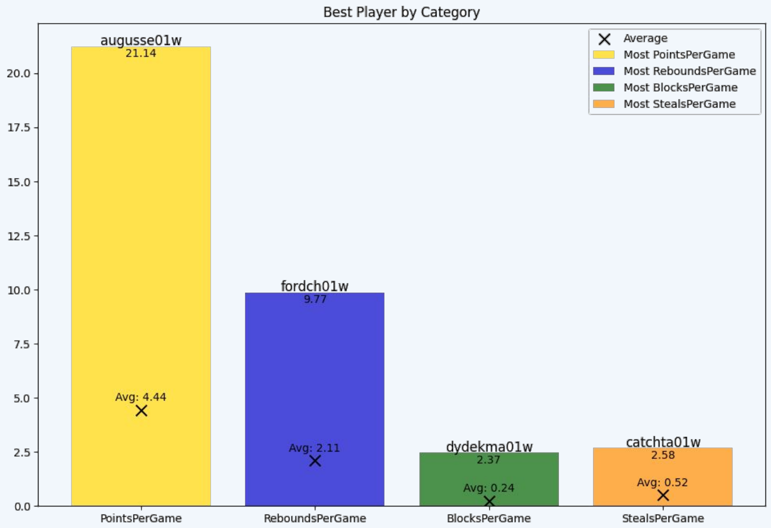


Fig 4 - Best Players by Category

Exploratory Data Analysis

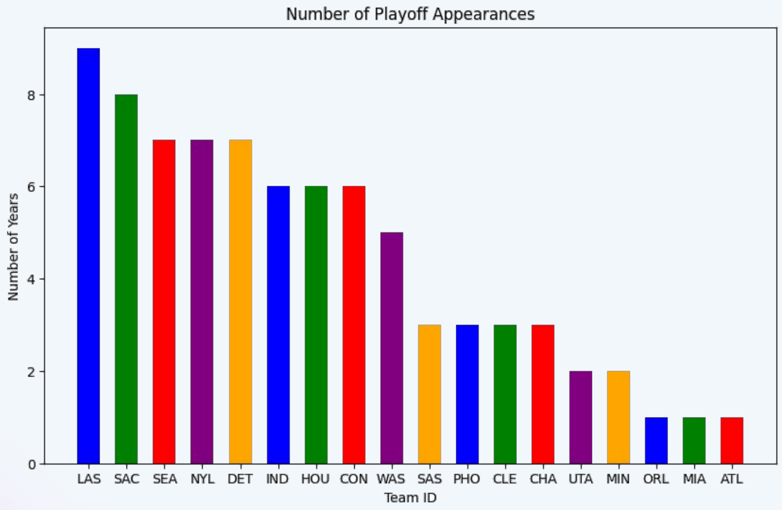


Fig 5 - Teams Playoff Appearances

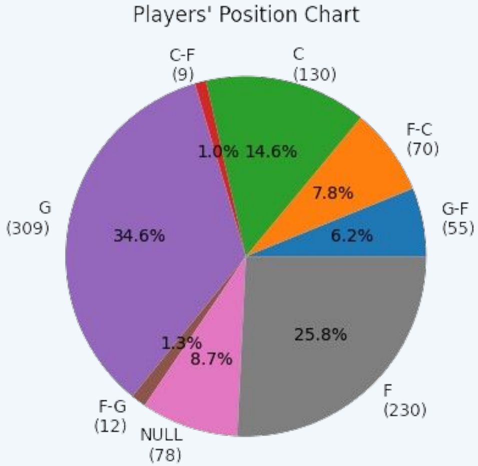


Fig 6 - Players Position Distribution

Predictive CRISP-DM Problem Definition

Task

- **Supervised Learning** project (Classification).
- **Binary** Target.
- **Predict** whether a team will qualify for the playoffs.
- **Balanced** dataset.

Experience

- 10 Years of data from the WNBA
- To predict a **specific** year, we will use the data from the **previous years**.

Performance Measure

- We will be using mostly **Accuracy** & **AUC** to evaluate our model.
- Kaggle submission results.

Data Preparation

Coaches

+ Regular & Playoff win-rate;

+ Coach_Awards;

+ Num_Playoff appearances.

Teams

- Irrelevant attributes (arena, etc...);

+ Team Rating based on its players;

+ Replaced some features by its success rate; e.g made/attempted;

+ Team Power Ratings based on current Player Ratings;

+ Playoff Rank;

+ Binary Encoding Target Variable.

Players

- Pointless attributes, **always** the **same** value (e.g first & lastseason);

- Players without any played game. (338 entries)

+ **Low** Null % - Replace by its mean. (weight & height)

Outliers

- Box plots;
- Not many outliers found;
- Can't really be sure they are **real** outliers.

Players_Teams

+ Player_Awards;

+ Replaced some features by its success rate; e.g made/attempted;

+ **PER** attribute based on John Hollinger's Player Efficiency Rating formula;

+ Regular & Playoff Rating based on statistics.

+ Combined both Regular & Playoff stats (**weighted**)

Data Preparation (Ratings)

Feature Importance

• Random Forest feature importance to quantify the impact of each stat on whether a player makes the playoffs.

Regular/Playoff Rating

• Calculated a player rating for each phase of the season (Playoffs and Regular), taking in consideration his statistics and the importance of each.

Final Rating

• Combined both ratings together into a single attribute. (**Weighted**)

Team Rating

- Teams have **two** ratings;
- One taking into consideration the players are playing the **current** season;
- The other the players are played the **last** season; (Speaks to if the team normally has good players)

Feature Selection

Correlation Matrix

- Checked for correlation between the continuous attributes;
- Removed highly correlated features.

Point-Biserial Correlation

- Removed continuous features with very **low** correlation to the target.

Principal Component Analysis

- Reduces dimensionality by finding new uncorrelated variables.

Recursive Feature Elimination

- Iteratively removes features to see which features generate the best results for each model.

Force Model Output

- Since only **eight** teams actually qualify for the playoffs, we can force the model to predict only eight **1's**. (The 4 highest probabilities of each conference)

Experimental Setup

Data Preparation & Pre-Processing

Merging data into a **single** dataframe

Data **Normalization**

Feature Selection

Testing different **Classification** Models (Random Forest, Logistic Regression, Gradient Boosting, SVM & KNN)

Hyperparameters of each model are tuned using **GridSearch**

Analysing the results, putting more emphasis on the **Accuracy** & **AUC**

Selecting the best **setting** to test the submission dataset

Models Performance

- **Average Performance Results**
- Each **YEAR** is trained with the **previous years**

Classification Model	Accuracy	AUC
Logistic Regression	0.71	0.69
SVM	0.69	0.69
Random Forest	0.64	0.64
Gradient Boosting	0.60	0.58
KNN	0.56	0.53

Models Performance

- **Average Performance Results**
- Each **YEAR** is trained with the **previous years**

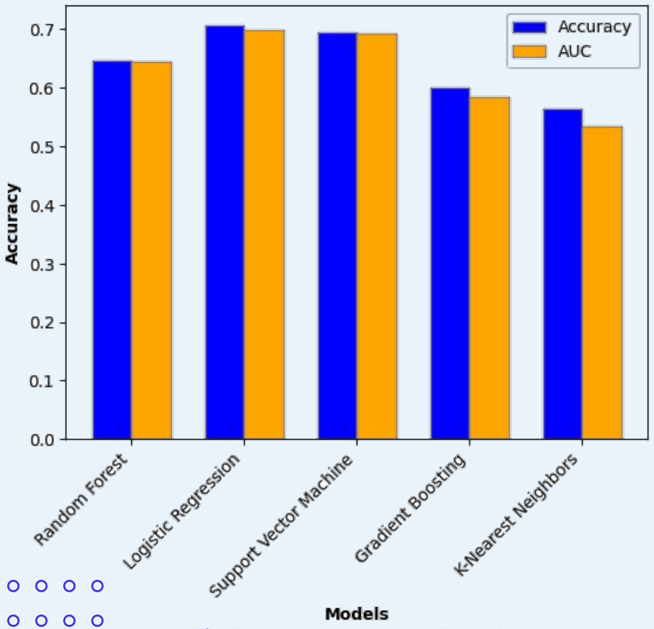


Fig 7 - Average Model Results

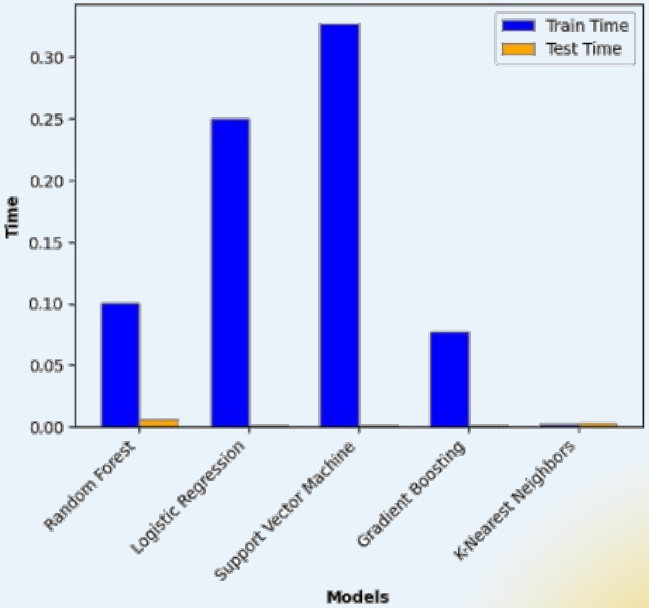


Fig 8 - Average Model Time

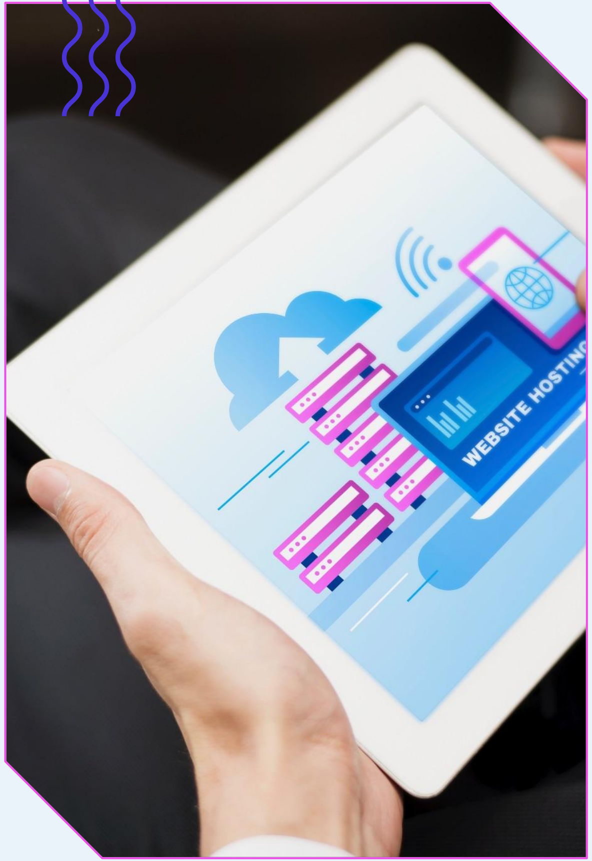
Kaggle Test Year

Submission

TeamID	Playoff
ATL	Y
CHI	N
CON	N
IND	Y
LAS	Y
MIN	Y
NYL	Y
PHO	Y
SAS	Y
SEA	Y
TUL	N
WAS	Y

Conclusions

- The process of understanding and exploring the data was very important in the early phases.
- The data cleaning and preparation made sure that the data was ready for modeling.
- We noted different results in model performance, with Random Forest and Gradient Boosting performing the best.
- The role of feature selection was crucial in improving model results.
- This process involved continuous loops of understanding, preparing, modeling, and refining. **(CRISP-DM)**
- The results ensured confidence in our ability to accurately predict the teams that will make the playoffs.



05

Annexes

Annexes

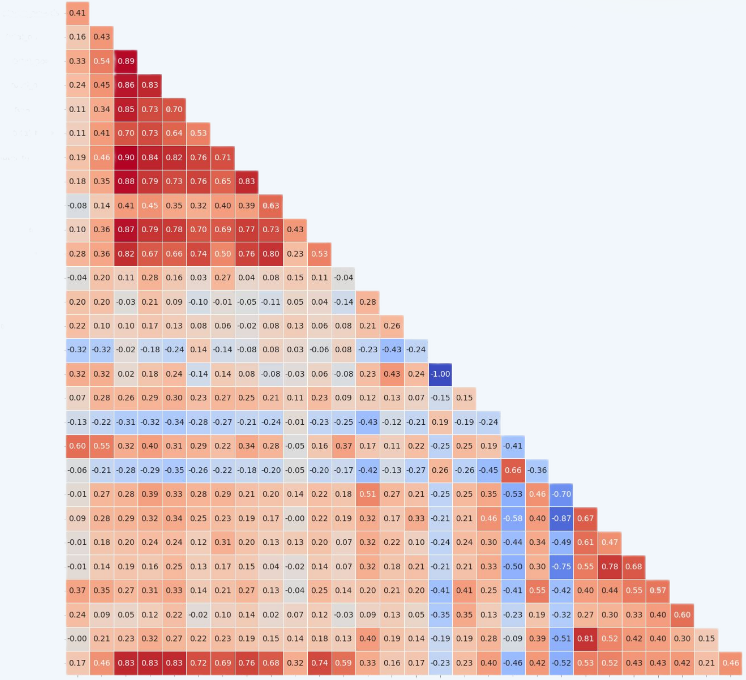


Fig 10 - Correlation Matrix

HOW TO CALCULATE PER:

[FGM x 85.910
+ Steals x 53.897
+ 3PTM x 51.757
+ FTM x 46.845
+ Blocks x 39.190
+ Offensive_Reb x 39.190
+ Assists x 34.677
+ Defensive_Reb x 14.707
- Foul x 17.174
- FT_Miss x 20.091
- FG_Miss x 39.190
- TO x 53.897]
x (1 / Minutes)

PER Formula, used to calculate the player rating

Correlation Matrix

Pairs > **0.75** correlation

total_minutes and **total_points**: 0.89
total_minutes and **total_steals**: 0.85
total_minutes and **total_turnovers**: 0.90
total_minutes and **total_pf**: 0.88
total_points and **total_turnovers**: 0.84
total_assists and **total_minutes**: 0.86
total_assists and **total_points**: 0.83
total_assists and **total_turnovers**: 0.82
total_assists and **total_gs**: 0.78
total_steals and **total_turnovers**: 0.76
total_pf and **total_points**: 0.79
total_pf and **total_steals**: 0.76
total_pf and **total_turnovers**: 0.83
total_gs and **total_minutes**: 0.87
total_gs and **total_points**: 0.79
total_gs and **total_turnovers**: 0.77
total_gp and **total_minutes:** 0.82
total_gp and **total_turnovers**: 0.76
total_gp and **total_pf**: 0.80
total_drebounds_pct and **total_orebounds_pct**: -1.00
playoff_rank and **po_winrate**: -0.87
coach_po_wr and **playoff_rank**: -0.75
coach_po_wr and **po_winrate**: 0.78
team_rating and **winrate**: 0.81
team_players_rating and **total_minutes**: 0.83
team_players_rating and **total_points**: 0.83
team_players_rating and **total_assists**: 0.83
team_players_rating and **total_turnovers**: 0.76

Point Biserial Correlation Test

po_winrate: 45.62% correlation
playoff_rank: 45.04% correlation
team_players_rating: 39.27% correlation
winrate: 34.91% correlation
coach_po_wr: 32.54% correlation
total_assists: 29.99% correlation
coach_reg_wr: 29.64% correlation
total_points: 28.92% correlation
player_awards: 27.79% correlation
total_blocks: 26.97% correlation
total_minutes: 26.44% correlation
team_rating: 26.39% correlation
total_turnovers: 25.48% correlation
coach_playoffs_count: 24.50% correlation
rank: 24.16% correlation
total_gs: 23.08% correlation
total_steals: 22.88% correlation
total_pf: 21.06% correlation
team_playoffs_count: 19.10% correlation
total_drebounds_pct: 14.76% correlation
total_orebounds_pct: 14.76% correlation
total_ft_pct: 13.38% correlation
coach_awards: 12.93% correlation
total_fg_pct: 11.87% correlation
total_dq: 11.23% correlation
total_gp: 8.83% correlation
total_three_pct: 7.22% correlation

Annexes

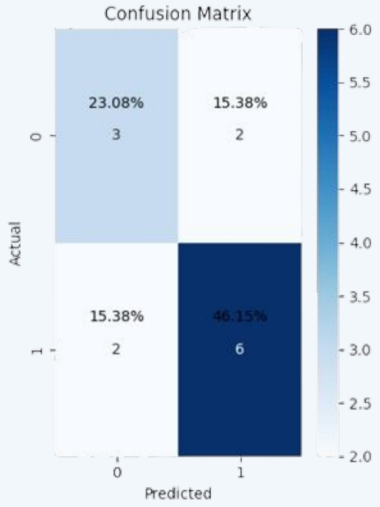


Fig 11 - Random Forest Year 10

	precision	recall	f1-score	support
0.0	0.60	0.60	0.60	5
1.0	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

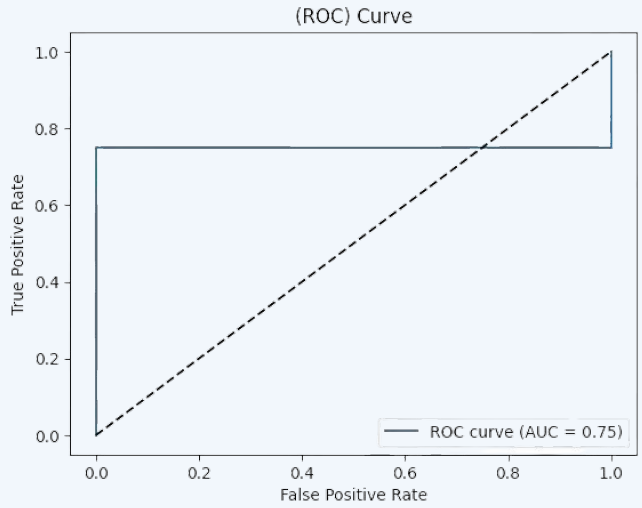
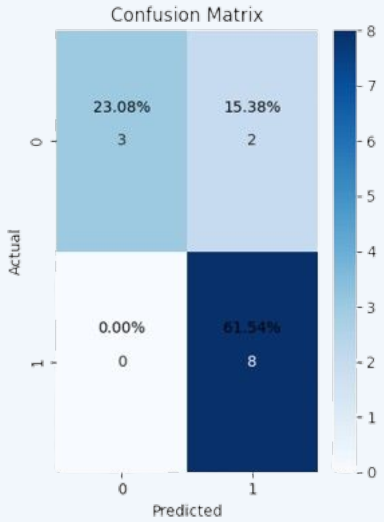


Fig 12- Random Forest ROC Curve Year 10

Annexes



	precision	recall	f1-score	support
0.0	1.00	0.60	0.75	5
1.0	0.80	1.00	0.89	8
accuracy			0.85	13
macro avg	0.90	0.80	0.82	13
weighted avg	0.88	0.85	0.84	13

Fig 13 - Logistic Regression Year 10

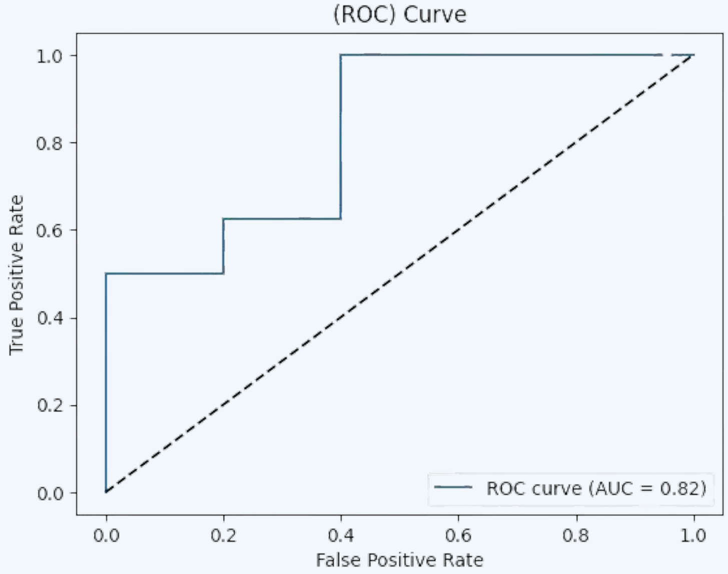
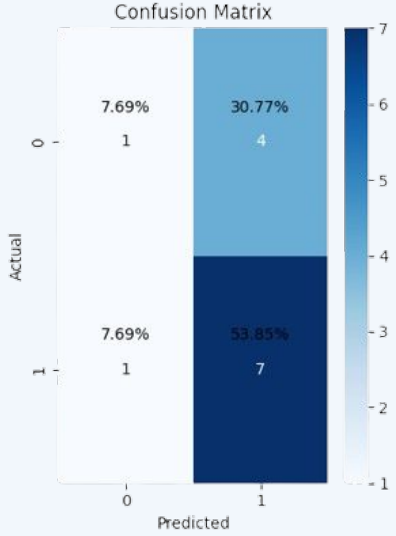


Fig 14 - Logistic Regression ROC Curve Year 10

Annexes



	precision	recall	f1-score	support
0.0	0.50	0.20	0.29	5
1.0	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

Fig 15 - SVM Year 10

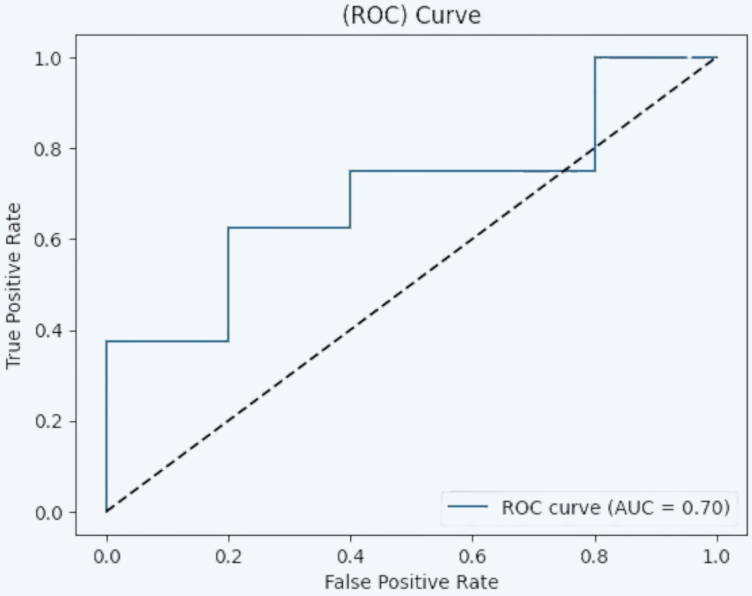


Fig 16 - SVM ROC Curve Year 10

Annexes

	precision	recall	f1-score	support
0.0	0.75	0.60	0.67	5
1.0	0.78	0.88	0.82	8
accuracy			0.77	13
macro avg	0.76	0.74	0.75	13
weighted avg	0.77	0.77	0.76	13

Fig 17 - Gradient Boosting Year 10

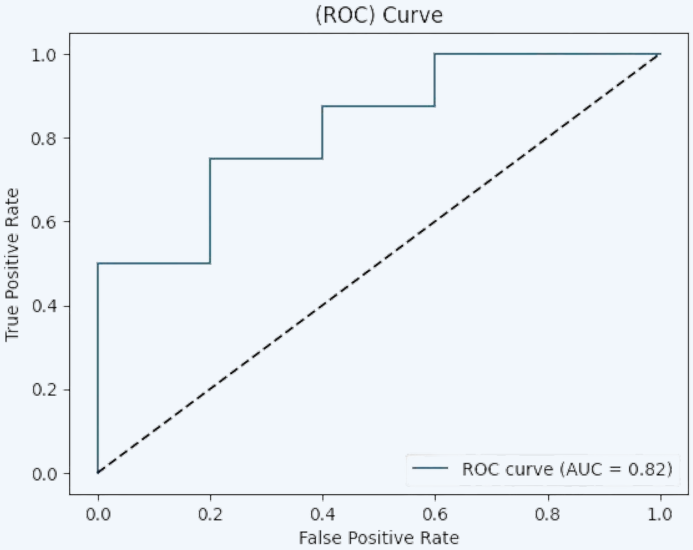


Fig 18 - Gradient Boosting ROC Curve Year 10

Annexes

Fig 19 - KNN Year 10

	precision	recall	f1-score	support
0.0	1.00	0.60	0.75	5
1.0	0.80	1.00	0.89	8
accuracy			0.85	13
macro avg	0.90	0.80	0.82	13
weighted avg	0.88	0.85	0.84	13

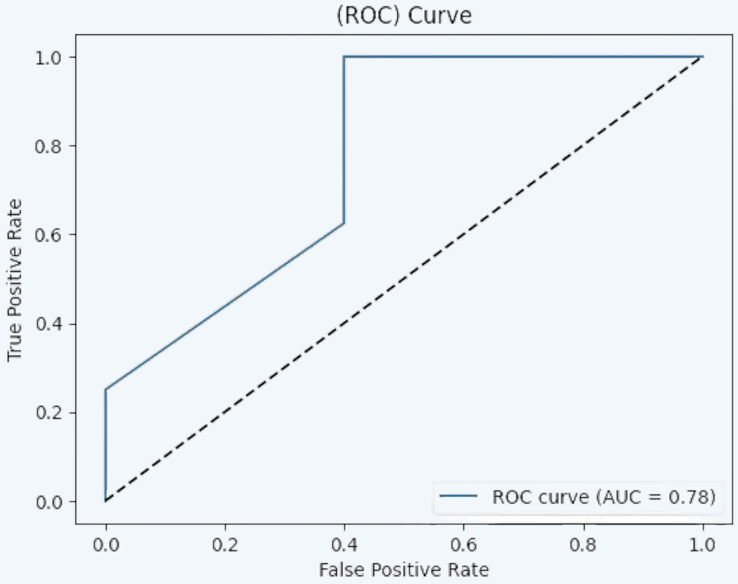


Fig 20 - KNN ROC Curve Year 10