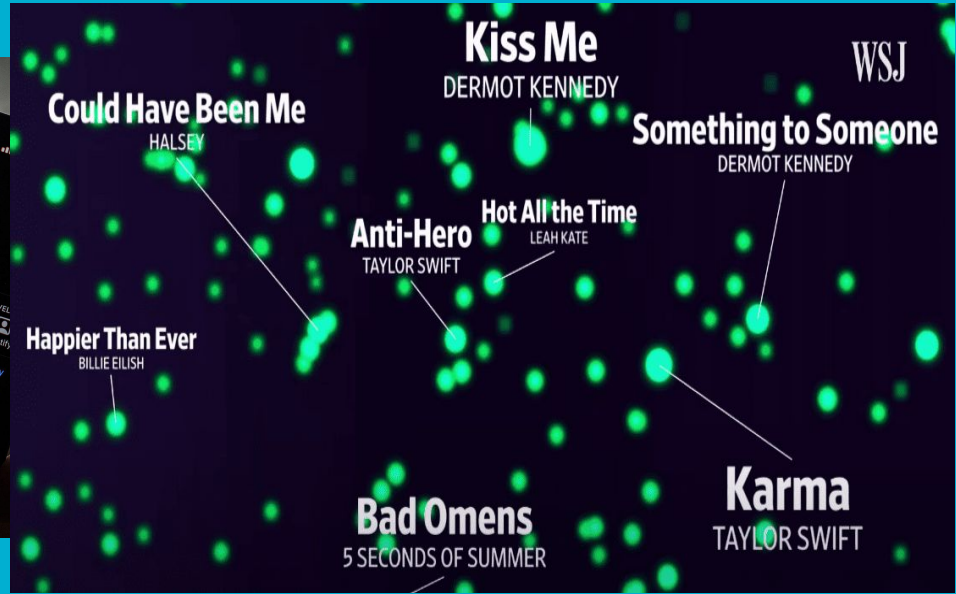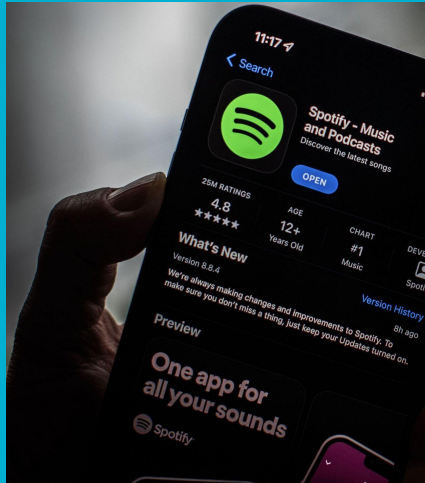# Machine Learning mini project

—

Nicolás Muñoz

Project Overview.

- Dataset: **Football Players Data**

- **Description:**

This dataset consists of 17.954 rows and 51 columns.

It emcompases a wide array of players data points, including players name, nationality, positions, player ratings, potential, and various skill attributes.

Source: Kaggle.com

Machine Learning mini Project

The aim of this mini project is to make predictions on Football player's market value, using their most significant attributes.
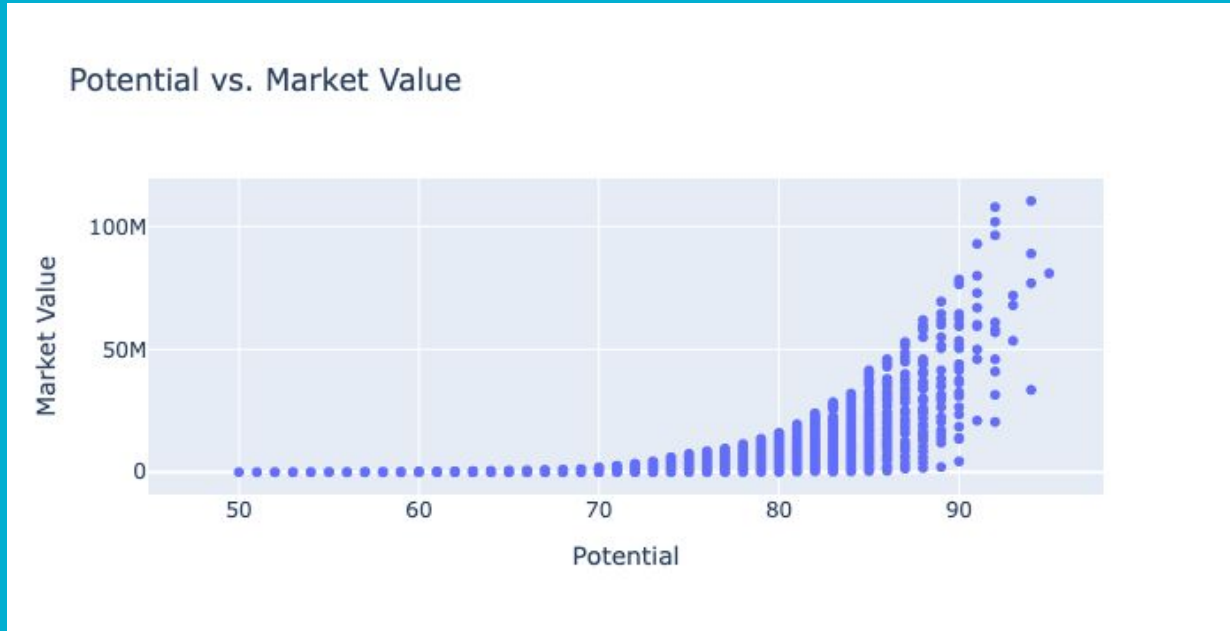
Data Selection and Preparation

- Dropping Missing Values and Columns
- Transforming columns such as ['positions'] and ['preferred_foot'], applying **lambda** function.

- Categorizing columns such as ['positions'], applying **map** function .

	age	height_cm	weight_kgs	positions	overall_rating	potential	value_euro	wage_euro	preferred_foot	weak_foot(1-5)	skill_moves(1-5)	national_rating
0	31	170.18	72.1	attack	94	94	110500000.0	565000.0	1	4	4	82.0
1	27	154.94	76.2	midfield	88	89	69500000.0	205000.0	0	5	4	78.0
2	25	190.50	83.9	midfield	88	91	73000000.0	255000.0	0	4	5	84.0
3	27	162.56	59.0	attack	88	88	62000000.0	165000.0	0	4	4	83.0
5	27	193.04	92.1	defence	88	90	59500000.0	215000.0	0	3	2	81.0
...	...	...	...	...	...	...	...	...	...	...	...	...
17940	28	172.72	76.2	attack	91	91	93000000.0	340000.0	0	4	4	85.0
17941	27	154.94	69.9	midfield	91	92	102000000.0	355000.0	0	5	4	85.0
17942	28	193.04	76.2	GK	91	93	72000000.0	260000.0	0	3	1	85.0
17943	27	175.26	68.0	attack	92	92	108000000.0	290000.0	0	5	5	81.0
17944	34	187.96	83.0	attack	94	94	77000000.0	405000.0	0	4	5	82.0

857 rows × 41 columns

Feature Engineering and Selection.

In Feature engineering for a **regression problem** is common to handle numerical and categorical data differently (because they have different characteristics).

- Grouping numerical data
- Grouping categorical data

The categorical features need conversion into a numerical format. Techniques like **one-hot encoding** are utilized to transform categories into numbers.

```
categorical_num= pd.get_dummies(categorical, drop_first = True)
categorical_num
```

|	positions_attack	positions_defence	positions_midfield
0	True	False	False
1	False	False	True
2	False	False	True
3	True	False	False
5	False	True	False

Model building and evaluation

- **Key Features**: Includes player ages, positions, overall rating, potential, financial information and skills.

  Target: The focus is on value_euro, representing each player market value.

- **Train-test Split:** The data was split into an 80-20 training and testing set.

Feature Scaling:

Box Plot of Normalized Features
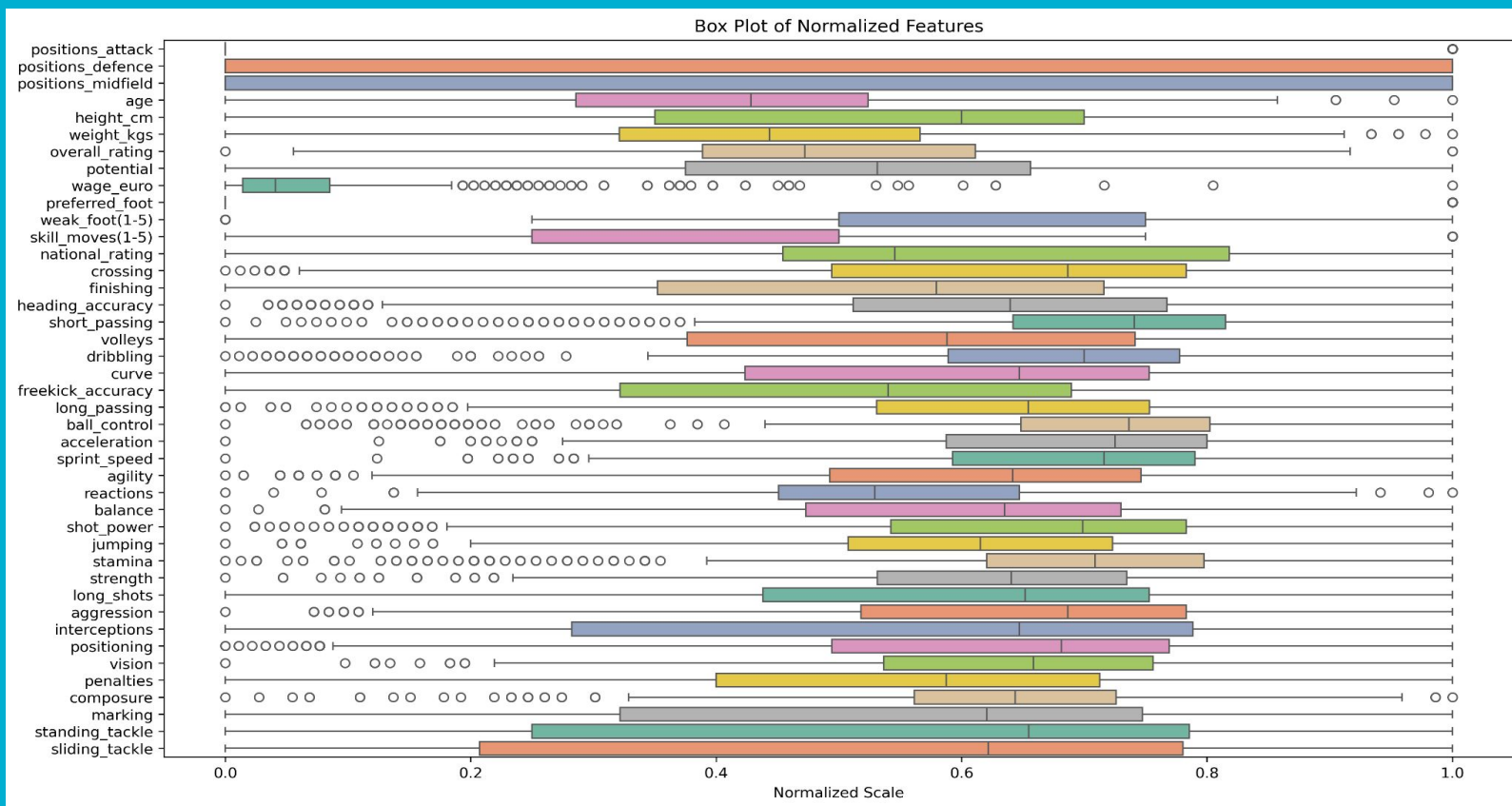
Machine Learning Models used:

1	**Knn Regressor**	• Knn score: 0.67
1	**Linear Regression**	• MAE: 4809558.736 • RMSE: 7954569.17 • R2 score: 0.80
3	**Decision Tree**	• MAE: 3673120.83 • RMSE: 6529944.26 • R2 score: 0.86
4	**Random Forest**	• MAE: 1602078.77 • RMSE: 4343648.07 • R2 score: 0.94
4	**AdaBoost**	• MAE: 1681017.44 • RMSE: 4429201.77 • R2 score: 0.93

Decision Tree evaluation
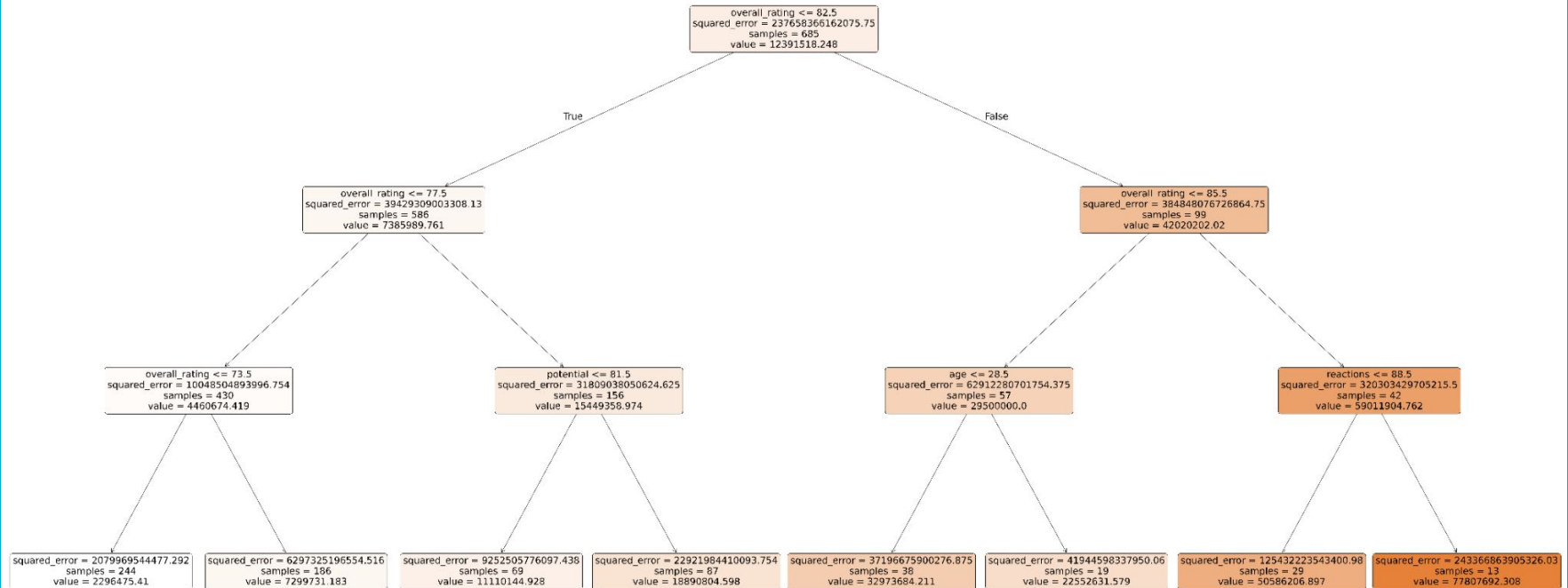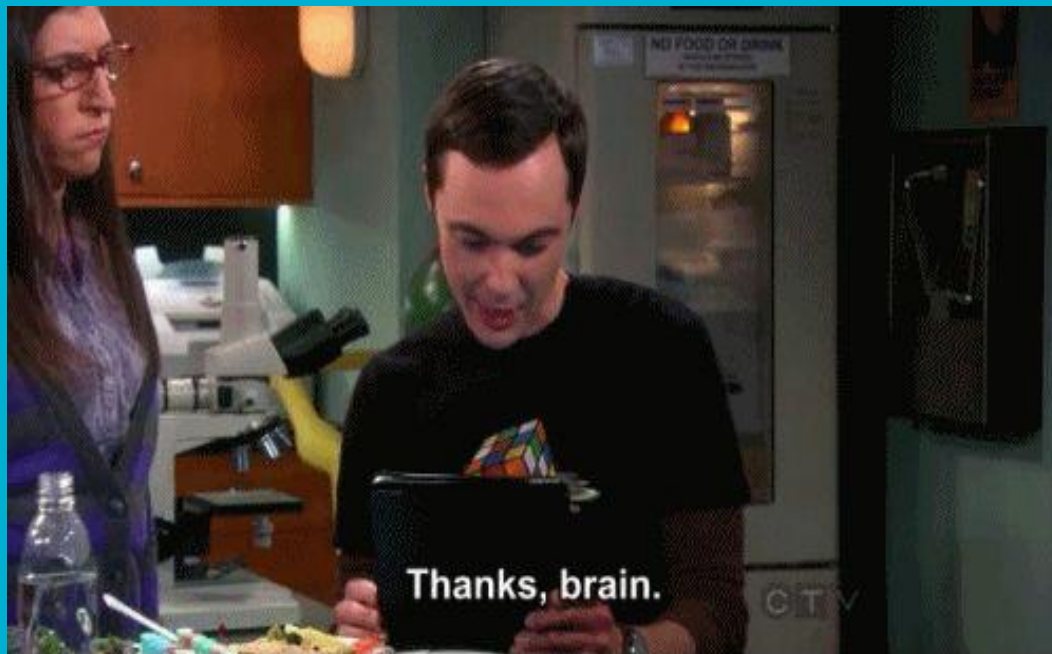
Challenges and Learnings:

- Presentations skills
- Integration of new content in a fast paced.

Future work and improvement:

- Better practices and techniques to work with numerical values in the feature engineering preprocessing.

Machine Learning mini project

Thanks!

Nicolás Muñoz