**DATA SCIENCE PROJECT REPORT ON FIFA20**

**Project Team ID:** PTID-CDS-AUG-23-1607B

**Project ID:** PRCP-1004-Fifa20

**BUSINESS CASE: BASED ON THE FEATURES OF DATA WE NEED TO CLUSTER (GROUP) THE PLAYER BASE ON THEIR SKILLSET**

**Abstract:**

The FIFA 20 project aims to develop a robust player clustering model using machine learning techniques. The primary objective is to create a tool that can effectively group football players based on their skillsets and attributes.

Methodology involves data preprocessing, exploratory data analysis (EDA), feature engineering, and the application of machine learning algorithms. We used clustering techniques such as K-means clustering to group players based on their playing style, skills, and other relevant attributes.

**Device Project In-to Multiple Steps:**

1. Data Collection
2. Loading data
3. Domain Analysis
4. .Basic Checks of data
5. EDA (Univariate, Bivariate, Multivariate Analysis)
6. Data Pre-processing
7. Feature Selection
8. Building ML Model
9. Training & Model Evaluation
10. Model Savings

**Data Collection:**

The dataset received from the Datamites educational platform and provided in a spreadsheet format, specifically as a CSV File.The CSV file contains the necessary fields and columns related to automobiles, including features such as make, Engine type, fuel type, and other relevant specifications.

**Loading data:**

load data in python using pandas library

**DOMAIN ANALYSIS**

Understanding the meaning of each feature and understanding the significance of each attribute in defining a player's skillset is essential.

**EDA (Univariate, Bivariate, Multivariate Analysis)**

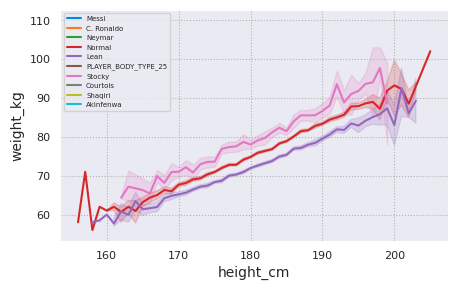
1. **Univariate Data Analysis**

Use sweetviz library and generate a html report of all feature to do univariate analysis, in that we get the Minimum, Maximum, Some statistical information of the particular feature.

1. **Bivariate Data Analysis**

In Bivariate analysis we check the relation of independent features to each other.

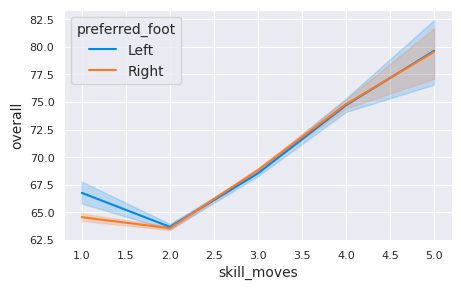
1. **Multivariate Data Analysis¶**
2. **Relation Between Height & weight with respect to Body Type.**

****

**Observation/Insight**

In this graph as height is increasing weight is also increasing with respect to three types of body normal, lean, Courtois.

1. **Relation between skill moves and overall with respect to preferred foot**

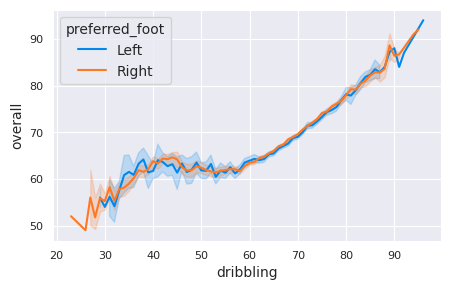
****

**Observation/insight**

the overall skill rating for the score of two is very low with respect to both foot.

on the other hand skill\_moves increasing range from 4 to 5 along with their overall skill rating is also increasing with both right and left foot.

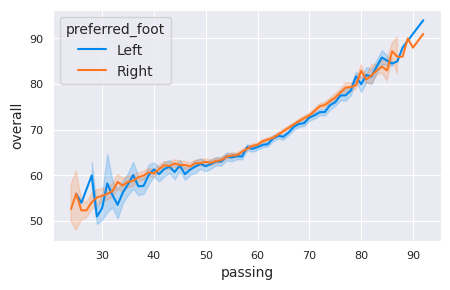
1. **Relation between dribbling and overall with respect to preferred foot**



**Observation/Insights**

• In this graph we see the relationship between dribbling and overall with respect to preferred\_foot and we clearly seen that dribbling skill is increasing at the same time overall skill rating also increasing with respect to preferred right or left foot.

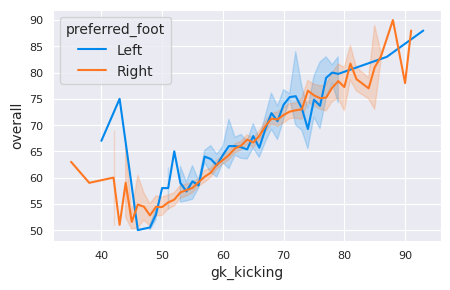
1. **Relation between Passing and overall with respect to preferred foot**

****

**Observation/Insights**

• In this graph we see the relationship between passing and overall with respect to preferred\_foot here clearly seen that passing score increased at the same time overall skill rating are also increase with both preferred foot.

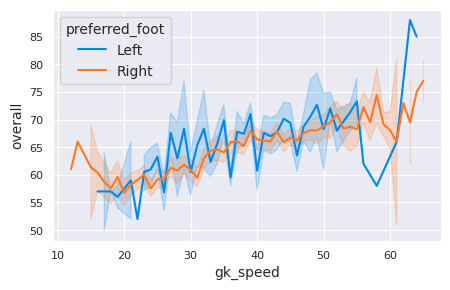
1. **Relation between Gk\_Kicking and overall with respect to preferred foot**

****

**Observation/Insights**

• preferred right and left foot at the start are both slightly high as compared to where the gk\_kicking is increasing along with their overall skill rating is also increasing with both right and left foot.

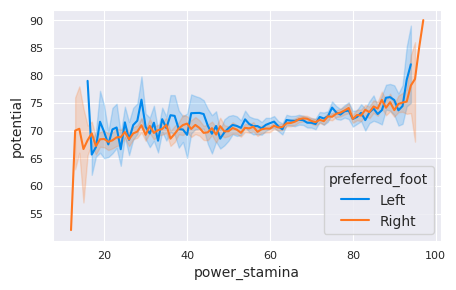
1. **Relation between Gk\_Speed and overall with respect to preferred foot**

****

**Observation/Insights**

In this graph preferred left foot with respect to their gk speed and overall skill rating is suddenly increase and suddenly decrease. but on the other hand preferred right foot with respect to gk speed and overall skill rating both are increasing.

1. **Relation between Power Stamina and Potential with respect to preferred foot**

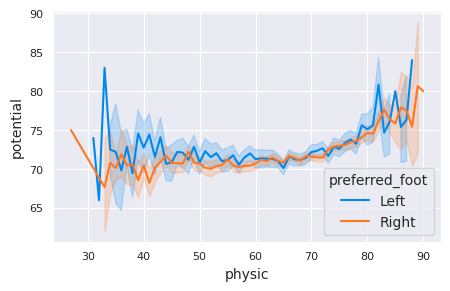
****

**Observation/Insights**

• At the start stages players with a preference for their right foot tend to have a low potential score. their power\_stamina is minimum a score of before 20, there is a suddenly drop in their potential rating. currently players with a preference for their left foot see a sudden increase in their potential rating it rises to 80.

on the other hand power\_stamina increasing along with their score of potential is also increasing with both right and left foot.

1. **Relation between Physic and Potential with respect to preferred foot**



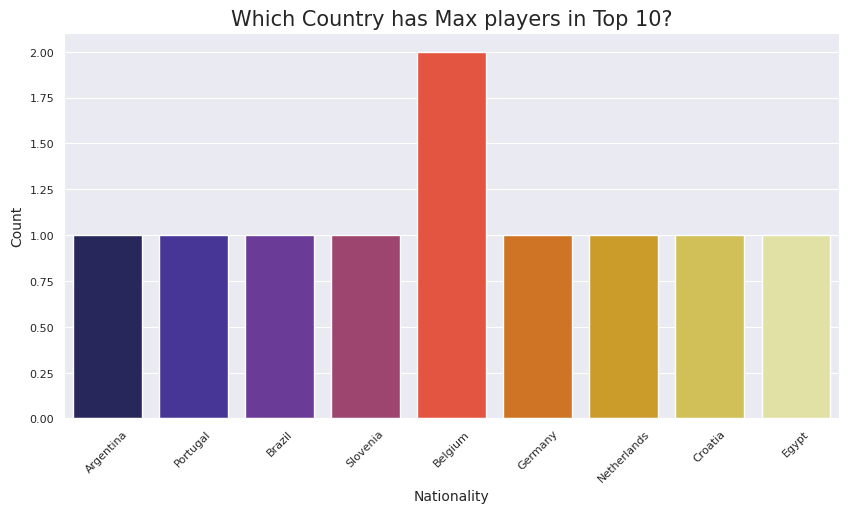
**Observation/Insights**

preferred left foot has gone slightly above 80 score of potential at the start and the preferred right foot has gone up to 75 score of potential at the start.

physic is increasing along with their score of potential is also increasing with both right and left foot.

**Here Some Condition & Plotting**

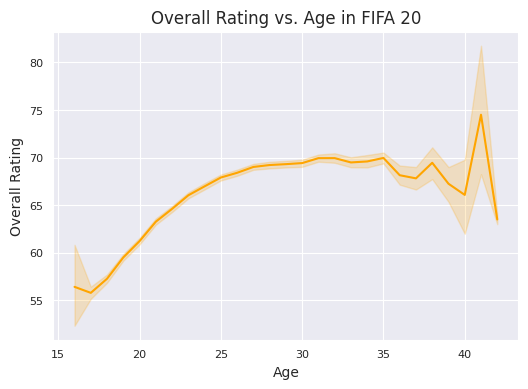
**1.prepare a rank ordered list of top 10 countries with most players. Which countries are producing the most footballers that play at this level?**



**Observation/Insights**

Belgium country has the highest representation player among the top 10.

**2.plot the distribution of overall rating vs. Age of players. Interpret what is the age after which a player stops improving?**



**Observation/Insights**

It can be estimated that a player typically stops improving after the age of 40.

**3.which type of offensive player tends to get paid the most: the striker, the right winger, the left winger?**

## 

**Observation/Insights**

The graph shows that left wingers have a higher salary, increasing 20,000, compared to both right wingers and strikers. Right wingers also tend to have relatively high salaries when compared to strikers.

**Data Preprocessing**

First we check the missing values, and then check missing values in percentage, we seen that the above 50% to 90% missing value and some unique feature also contain missing value so we drop this feature.

Remaining feature missing value is less than 50% so we impute the missing value with median and mode.

Second we Handle categorical data and use Manual encoding and frequency encoding. Because features has contain lots of label.

In this data I’m Clearly seen that some feature has lots of outlier & we impute them, for that first we check the distribution of all feature and plot the box plot and decide the technique. In this data we are handle only important feature outlier, because the remaining feature is unique or some feature is not required to handle outlier.

Scale the numerical independent feature with the help of Minmax scalar and scale the feature. Use min max scaling because of dataset contain large amount of outlier so outlier is going to be biased.

**Feature Scaling**

First we drop unique and constant feature Here we are going to drop unique column as well as lots of missing value column. The column ls,st,rs,lw,cf etc.. are playing position in the game and the data in this columns is basically the potential of the player if were to play in that position, so we assume the player only plays with the team position and we will drop this column.

Check the correlation with the help of heatmap and seen the From the above heatmap is very difficult to find highly correlated feature so we are create a python code to check the highly corelated feature and drop highly correlated feature.

The dataset not contain any duplicates.

After that save & load the preprocess data, Save the dataframe to a CSV file.

Using PCA(PRINCIPLE COMPONENET ANALYSIS) to reduce the feature. Here we are select 10 components because less variance loss

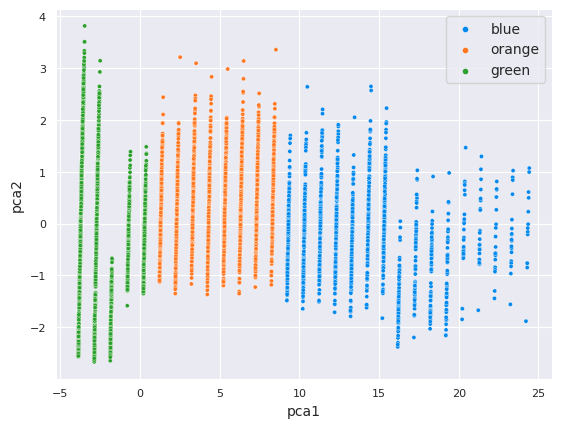
**Model Creation & Evaluation**

Define Independent variable.

Use K-Means Clustering algorithm using elbow Method technique to determining the optimal number of clusters (k) in a K-means clustering algorithm. From the plot we are select 3 cluster because of odd no and more variance.

Using K-Means Clustering to get a Silhouette Score is 0.62.

**Cluster**



**Model Saving**

Save the model using pickle file.

**Conclusion**

In this FIFA20 data science project, we embarked on a journey to create a player clustering model based on their skillsets and attributes. Through a systematic approach encompassing data collection, exploratory data analysis (EDA), data preprocessing, and clustering, we gained valuable insights into the virtual footballing world. We successfully applied the K-Means clustering algorithm and determined that three clusters provide a meaningful grouping of players. This clustering method allows us to categorize players into distinct segments based on their skillsets. Through multivariate data analysis, we observed several intriguing relationships between player attributes. For instance, players who exhibit a preference for their right foot tend to have distinct patterns in various attributes, including dribbling, passing, and potential. Belgium emerged as the country with the highest representation of players in the top 10 countries. This finding sheds light on the diversity of nationalities in the FIFA20 player database. Our analysis suggests that players typically stop improving after the age of 40. This insight can be valuable for player development strategies in the game. Left wingers tend to have higher salaries compared to right wingers and strikers, indicating that this offensive player type is compensated more generously in the virtual football world. We meticulously handled missing values, outliers, and categorical data, ensuring the robustness of our clustering model. Feature scaling and dimensionality reduction through PCA enhanced the model's performance. The clustering model can provide valuable market insights by identifying trends in player attributes and preferences. This information can guide decisions related to player pricing and availability in the virtual transfer market.