**Predicting players rating**

In this project you are going to predict the overall rating of soccer player based on their

attributes such as 'crossing', 'finishing etc.

The dataset you are going to use is from European Soccer Database

(https://www.kaggle.com/hugomathien/soccer) has more than 25,000 matches and more

than 10,000 players for European professional soccer seasons from 2008 to 2016.

Download the data in the same folder and run the following commmand to get it in the environment.

**About the Dataset**

**The ultimate Soccer database for data analysis and machine learning**

The dataset comes in the form of an SQL database and contains statistics of about 25,000 football matches, from the top football league of 11 European Countries. It covers seasons from 2008 to 2016 and contains match statistics (i.e: scores, corners, fouls etc...) as well as the team

formations, with player names and a pair of coordinates to indicate their position on the pitch.

* +25,000 matches
* +10,000 players
* 11 European Countries with their lead championship
* Seasons 2008 to 2016
* Players and Teams' attributes\* sourced from EA Sports' FIFA video game series, including
* the weekly updates
* Team line up with squad formation (X, Y coordinates)
* Betting odds from up to 10 providers
* Detailed match events (goal types, possession, corner, cross, fouls, cards etc...) for
* +10,000 matches

The dataset also has a set of about 35 statistics for each player, derived from EA Sports' FIFA video games. It is not just the stats that come with a new version of the game but also the weekly updates. So for instance if a player has performed poorly over a period of time and his stats get

impacted in FIFA, you would normally see the same in the dataset.

**Python skills required to complete this project**

**SQL:**

The data is in SQL database so students need to retrive using query language. They also need to know how to connect SQL database woth python. The library we are using for this in 'sqlite3'.

SQLite3 can be integrated with Python using sqlite3 module, which was written by Gerhard Haring. It provides an SQL interface compliant with the DB-API 2.0 specification described by PEP 249. You do not need to install this module separately because it is shipped by default along with Python version 2.5.x onwards. To use sqlite3 module, you must first create a connection object that represents the database and then optionally you can create a cursor object, which will help you in executing all the SQL

statements.

**Pandas:**

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including

finance, economics, Statistics, analytics, etc.In this tutorial, we will learn the various features of Python Pandas and how to use them in practice.

**Scikit Learn**

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

* NumPy: Base n-dimensional array package
* SciPy: Fundamental library for scientific computing
* Matplotlib: Comprehensive 2D/3D plotting
* IPython: Enhanced interactive console
* Sympy: Symbolic mathematics
* Pandas: Data structures and analysis

Extensions or modules for SciPy care conventionally named SciKits. As such, the module provides learning algorithms and is named scikit-learn.

The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as easy of use, code quality, collaboration, documentation and performance.

**Machine Learning skills required to complete the project**

**Supervised learning**

Supervised learning deals with learning a function from available training data. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

**Regression**

Regression is a parametric technique used to predict continuous (dependent) variable given a set of independent variables. It is parametric in nature because it makes certain assumptions (discussed next) based on the data set. If the data set follows those assumptions, regression gives

incredible results.

**Model evaluation**

Student must know how to judge a model on unseen data. What metric to select to judge the performance

Let's get started.....

**Import Libraries**

import sqlite3

import pandas as pd

from sklearn.tree import DecisionTreeRegressor from

sklearn.linear\_model import LinearRegression from

sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error from

math import sqrt

**Read Data from the Database into pandas**

# Create your connection.

cnx = sqlite3.connect('database.sqlite')

df = pd.read\_sql\_query("SELECT \* FROM Player\_Attributes", cnx)

df.head()

**Task:**Deploy this assignment in any cloud platform.(Try to look for free cloud platform)

**Assignment:** Submit assignment’s deployable link only.

**Import Libraries**

**!**pip install xgboost

**import** sqlite3

**import** numpy **as** np

**import** pandas **as** pd

**%matplotlib** notebook

**import** matplotlib.pyplot **as** plt

**import** xgboost **as** xgb

**from** xgboost.sklearn **import** XGBRegressor

**from** xgboost **import** plot\_importance

**from** sklearn.linear\_model **import** LinearRegression

**from** sklearn.tree **import** DecisionTreeRegressor

**from** sklearn.ensemble **import** RandomForestRegressor

**from** sklearn.preprocessing **import** StandardScaler

**from** sklearn.impute **import** SimpleImputer **as** Imputer

**from** sklearn.feature\_selection **import** SelectFromModel

**from** sklearn.model\_selection **import** train\_test\_split, GridSearchCV, ShuffleSplit, RandomizedSearchCV

**from** sklearn.pipeline **import** make\_pipeline

**import** pickle

cnx **=** sqlite3**.**connect('database.sqlite')

dd **=** pd**.**read\_sql\_query("SELECT name FROM sqlite\_master WHERE type='table'", cnx)

print(dd)

df **=** pd**.**read\_sql\_query("SELECT \* FROM Player\_Attributes", cnx)

df**.**head()

target **=** df**.**pop('overall\_rating')

type(target)

df**.**shape

target**.**head()

## Imputing target funtion :

target**.**isnull()**.**values**.**sum()

target**.**describe()

y **=** target**.**fillna(target**.**mean())

y**.**isnull()**.**values**.**any()

## Data Exploration :

df**.**columns

**for** col **in** df**.**columns:

unique\_cat **=** len(df[col]**.**unique())

print("{col}--> {unique\_cat}..{typ}"**.**format(col**=**col, unique\_cat**=**unique\_cat, typ**=**df[col]**.**dtype))

dummy\_df **=** pd**.**get\_dummies(df, columns**=**['preferred\_foot', 'attacking\_work\_rate', 'defensive\_work\_rate'])

dummy\_df**.**head()

X **=** dummy\_df**.**drop(['id', 'date'], axis**=**1)

## Feature selection :

X\_train, X\_test, y\_train, y\_test **=** train\_test\_split(X, y, test\_size**=**0.25,

random\_state**=**42)

*#imputing null value of each column with the mean of that column*

imput **=** Imputer()

X\_train **=** imput**.**fit\_transform(X\_train)

X\_test **=** imput**.**fit\_transform(X\_test)

*#finding feature\_importance for feature selection. from it we'll be able to decide threshold value*

model **=** XGBRegressor()

model**.**fit(X\_train, y\_train)

print(model**.**feature\_importances

selection **=** SelectFromModel(model, threshold**=**0.01, prefit**=True**)

select\_X\_train **=** selection**.**transform(X\_train)

select\_X\_test **=** selection**.**transform(X\_test)

select\_X\_train**.**shape

## Training different models :

### 1. Linear Regression :

pipe **=** make\_pipeline(StandardScaler(), LinearRegression())

param\_grid **=** {'linearregression\_\_n\_jobs': [**-**1]}

grid **=** GridSearchCV(pipe, param\_grid**=**param\_grid, cv**=**10)

grid**.**fit(select\_X\_train, y\_train)

grid**.**best\_params\_

lin\_reg **=** pickle**.**dumps(grid)

**2. Decision Tree :**

pipe **=** make\_pipeline(StandardScaler(), DecisionTreeRegressor(criterion**=**'mse', random\_state**=**0))

param\_grid **=** {'decisiontreeregressor\_\_max\_depth': [3, 5, 9, 13]}

grid **=** GridSearchCV(pipe, param\_grid**=**param\_grid, cv**=**10)

grid**.**fit(select\_X\_train, y\_train)

grid**.**best\_params\_

Dectree\_reg **=** pickle**.**dumps(grid)

**3. Ranom Forest :**

pipe **=** make\_pipeline(StandardScaler(),

RandomForestRegressor(n\_estimators**=**200, random\_state**=**0))

*#cv = ShuffleSplit(test\_size=0.2, random\_state=0)*

param\_grid **=** {'randomforestregressor\_\_max\_features':['sqrt', 'log2'],

'randomforestregressor\_\_max\_depth':[9, 10]}

grid **=** GridSearchCV(pipe, param\_grid**=**param\_grid, cv**=**5)

grid**.**fit(select\_X\_train, y\_train)

grid**.**best\_params\_

Randfor\_reg **=** pickle**.**dumps(grid)

**4. Xgboost regressor :**

pipe **=** make\_pipeline(StandardScaler(),

XGBRegressor(n\_estimators**=**200, random\_state**=**42))

*#cv = ShuffleSplit(n\_splits=10, random\_state=0)*

param\_grid **=** {'xgbregressor\_\_max\_depth': [5, 7],

'xgbregressor\_\_learning\_rate': [0.1, 0.3]}

grid **=** GridSearchCV(pipe, param\_grid**=**param\_grid, cv**=**5, n\_jobs**=** **-**1)

grid**.**fit(select\_X\_train, y\_train)

grid**.**best\_params\_

xgbreg **=** pickle**.**dumps(grid)

## Comparision between different models :

lin\_reg **=** pickle**.**loads(lin\_reg)

Dectree\_reg **=** pickle**.**loads(Dectree\_reg)

Randfor\_reg **=** pickle**.**loads(Randfor\_reg)

xgbreg **=** pickle**.**loads(xgbreg)

print("""Linear Regressor accuracy is {lin}

DecisionTree Regressor accuracy is {Dec}

RandomForest regressor accuracy is {ran}

XGBoost regressor accuracy is {xgb}"""**.**format(lin**=**lin\_reg**.**score(select\_X\_test, y\_test),

Dec**=**Dectree\_reg**.**score(select\_X\_test, y\_test),

ran**=**Randfor\_reg**.**score(select\_X\_test, y\_test),

xgb**=**xgbreg**.**score(select\_X\_test, y\_test)))