Soccer player predict

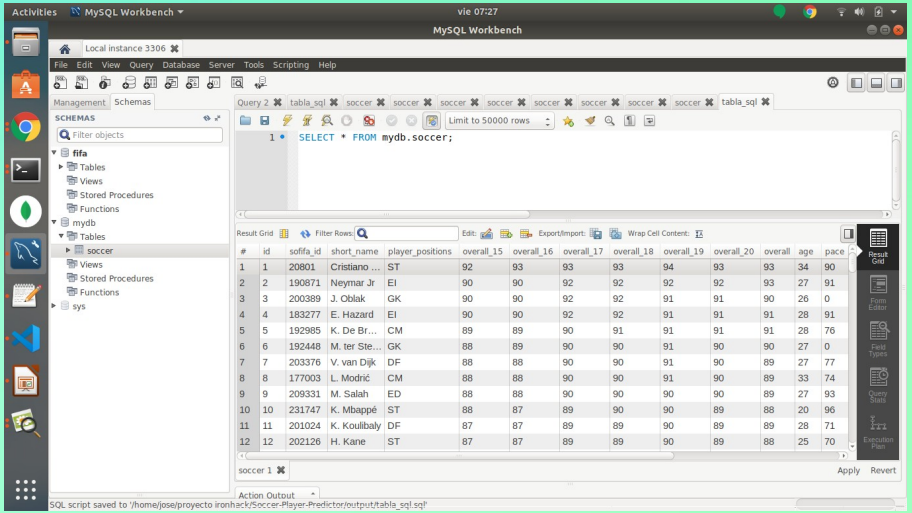
Aplicación para predecir la media de un jugador el año siguiente

Pasos seguidos

- Dataset de kaggle con datos de los futbolistas del juego FIFA
- Limpieza y filtrado
- Entrenar con varios modelos
- Predecir la progresion de cada jugador

Datasets

- Utilizo 6 datasets, correspondientes a las temporadas entre 2015 y 2019
- Contienen mas de 18000 jugadores y 103 parametros para cada uno



Entrenamiento:

Pruebo estos test:

- "Linear": LinearSVR(),
- "neigbor":KNeighborsRegressor(),
- "boosting": GradientBoostingRegressor(),
- "boosting2": GradientBoostingRegressor(n_estimators=500),
- "boosting3": GradientBoostingRegressor(n_estimators=1000),
- "histboosting": HistGradientBoostingRegressor(),
- "histboosting2": HistGradientBoostingRegressor(learning_rate=0.2),
- "histboosting3": HistGradientBoostingRegressor(learning_rate=0.4),
- "forest": RandomForestRegressor(),
- "forest250": RandomForestRegressor(n_estimators=250),
- "forest350": RandomForestRegressor(n_estimators=350),
- "tree": DecisionTreeClassifier(random_state=0, max_depth=2)

```
Training model: Linear
-17.683178842752802
Training model: neigbor
0.8704476517274042
Training model: boosting
0.975509605185762
Training model: boosting2
0.9875837222117004
Training model: boosting3
0.9902741499223167
Training model: histboosting
0.9903175510646335
Training model: histboosting2
0.9908268665093303
Training model: histboosting3
0.9899395717065198
Training model: forest
0.9927616522997671
Training model: forest250
0.9940337770529025
Training model: forest350
0.9940597949981074
Training model: tree
0.5307218242038364
```

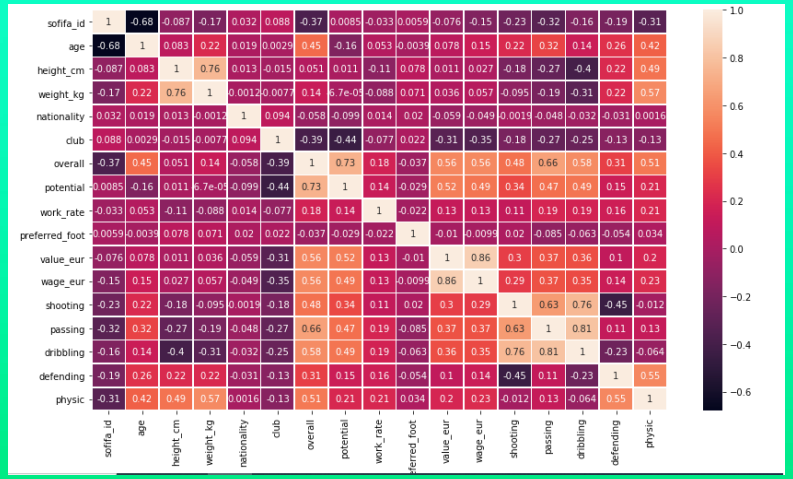
El ganador es: Training model: forest350 (0.9940597949981074)

Entrenamiento con H2O

model_id	rmse	mean_residual_deviance	mse	mae	rmsle
XGBoost_1_AutoML_20200309_161526	0.552025	0.304731	0.304731	0.373386	0.00896859
StackedEnsemble_AllModels_AutoML_20200309_161526	0.55303	0.305842	0.305842	0.37311	0.00898936
XGBoost_2_AutoML_20200309_161526	20.7582	430.902	430.902	19.813	0.375525

H2oAutoML buscará automáticamente modelos de machine learning que se ajustan a los datos y entrenaremos con el mejor

XGBoost_1_AutoML_20200308_215025Nos da el mejor resultado



Mapa de calor con la correlación de lascolumnas

```
In [11]: predictions = automl.leader.predict(h2test)

         xgboost prediction progress: |████████████████████████████████████████| 100%

In [13]: test['overall'] = predictions.as_data_frame()
         columnasentrega = ['sofifa_id','overall']
         entrega = test['overall']
         test = test[columnasentrega]
         test
```

Out[13]:

	sofifa_id	overall
0	158023	93.334595
1	20801	92.799469
2	190871	92.543037
3	200389	90.172699
4	183277	91.046577
...	...	...
18273	245006	49.266388
18274	250995	49.542862
18275	252332	49.423573
18276	251110	48.024727
18277	233449	49.274654

18278 rows × 2 columns

Tabla con las predicciones de la media del 2021

Gracias !!

Jose A. Vidal

sowe01@gmail.com
https://www.linkedin/in/jose-a-vidal/
https://github.com/sowe01