at-group4-week1-prepare-data

November 9, 2022

[1]:
```
import pandas as pd
import numpy as np

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

from joblib import dump
```

[2]:
```
df_2022_train = pd.read_csv('../data/raw/2022_train.csv')

df_main = df_2022_train.rename(columns = {
    'GP':'Games Played',
    'MIN':'Minutes Played',
    'PTS':'Points Per Game',
    'FGM' : 'Field Goals Made',
    'FGA' : 'Field Goals Attempts',
    'FG%' : 'Field Goals Percent',
    '3P Made' : '3Points Made',
    '3PA' : '3Points Attempts',
    '3P%' : '3Points Percent',
    'FTM' : 'Free Throw Made',
    'FTA' : 'Free Throw Attempts',
    'FT%' : 'Free Throw Percent',
    'OREB' : 'Offensive Rebounds',
    'DREB' : 'Defensive Rebounds',
    'REB' : 'Rebounds',
    'AST' : 'Assists',
    'STL' : 'Steals',
    'BLK' : 'Blocks',
    'TOV' : 'Turnovers'
    })

df_main.head()
```

[2]:

	Id	Games Played	Minutes Played	Points Per Game	Field Goals Made	\
0	3799	80	24.3	7.8	3.0	
1	3800	75	21.8	10.5	4.2	

```
2  3801            85            19.1              4.5              1.9
3  3802            63            19.1              8.2              3.5
4  3803            63            17.8              3.7              1.7

   Field Goals Attempts  Field Goals Percent  3Points Made  3Points Attempts  \
0                   6.4                 45.7           0.1               0.3
1                   7.9                 55.1          -0.3              -1.0
2                   4.5                 42.8           0.4               1.2
3                   6.7                 52.5           0.3               0.8
4                   3.4                 50.8           0.5               1.4

   3Points Percent  …  Free Throw Attempts  Free Throw Percent  \
0             22.6  …                 2.9                72.1
1             34.9  …                 3.6                67.8
2             34.3  …                 0.6                75.7
3             23.7  …                 1.5                66.9
4             13.7  …                 0.5                54.0

   Offensive Rebounds  Defensive Rebounds  Rebounds  Assists  Steals  Blocks  \
0                 2.2                 2.0       3.8      3.2     1.1     0.2
1                 3.6                 3.7       6.6      0.7     0.5     0.6
2                 0.6                 1.8       2.4      0.8     0.4     0.2
3                 0.8                 2.0       3.0      1.8     0.4     0.1
4                 2.4                 2.7       4.9      0.4     0.4     0.6

   Turnovers  TARGET_5Yrs
0        1.6            1
1        1.4            1
2        0.6            1
3        1.9            1
4        0.7            1

[5 rows x 21 columns]
```

[3]:
```
df_2022_test = pd.read_csv('../data/raw/2022_test.csv')

df_test = df_2022_test.rename(columns = {
    'GP':'Games Played',
    'MIN':'Minutes Played',
    'PTS':'Points Per Game',
    'FGM' : 'Field Goals Made',
    'FGA' : 'Field Goals Attempts',
    'FG%' : 'Field Goals Percent',
    '3P Made' : '3Points Made',
    '3PA' : '3Points Attempts',
    '3P%' : '3Points Percent',
    'FTM' : 'Free Throw Made',
```

```
    'FTA' : 'Free Throw Attempts',
    'FT%' : 'Free Throw Percent',
    'OREB' : 'Offensive Rebounds',
    'DREB' : 'Defensive Rebounds',
    'REB' : 'Rebounds',
    'AST' : 'Assists',
    'STL' : 'Steals',
    'BLK' : 'Blocks',
    'TOV' : 'Turnovers'
    })

df_test.head()
```

[3]:

	Id	Games Played	Minutes Played	Points Per Game	Field Goals Made	Field Goals Attempts	Field Goals Percent	3Points Made	3Points Attempts	3Points Percent	Free Throw Made	Free Throw Attempts	Free Throw Percent	Offensive Rebounds	Defensive Rebounds	Rebounds	Assists	Steals	Blocks	Turnovers
0	0	56	9.1	4.0	1.6	3.7	43.7	0.1	0.3	7.3	0.7	1.2	63.4	1.2	0.8	1.7	0.4	0.2	0.3	0.8
1	1	43	19.3	10.1	3.7	8.1	46.0	0.6	1.7	35.1	1.8	2.5	75.3	0.5	0.9	1.5	3.5	0.6	0.0	1.8
2	2	82	33.9	11.3	4.9	10.6	45.6	0.5	1.9	44.8	1.8	2.7	71.2	1.3	3.3	4.5	2.5	1.3	0.3	2.0
3	3	86	44.7	18.8	6.8	15.9	42.9	0.5	1.8	13.5	4.5	6.3	70.9	1.5	3.2	5.0	4.1	0.9	0.1	3.6
4	4	58	12.3	4.7	1.6	4.0	40.0	0.5	1.7	38.7	1.1	1.3	76.9	0.2	0.6	0.9	1.5	0.5	-0.4	0.9

[4]:
```
# Save data in the folder `data/processed`
np.save('../data/processed/alltrain', df_main)
np.save('../data/processed/test', df_test)
```

0.1 Standardise

Training Data

[5]:
```
df_standard = df_main.copy()
```

[6]:
```
target_var = df_standard.pop('TARGET_5Yrs')
target_var.head()
```

[6]:
```
0    1
1    1
2    1
3    1
4    1
Name: TARGET_5Yrs, dtype: int64
```

[7]:
```
scaler = StandardScaler()
df_standard = scaler.fit_transform(df_standard)

# Save the scaler into the folder models and call the file scaler.joblib
# using joblib - dump
dump(scaler, '../models/scaler.joblib')

# Save the different sets in the folder `data/processed`
np.save('../data/processed/df_standard', df_standard)
np.save('../data/processed/target_var', target_var)
```

[8]:
```
df_standard
```

[8]:
```
array([[-1.73183431,  1.00610018,  0.6405738 , …,  1.1072419 ,
        -0.05507101,  0.47321012],
       [-1.7314013 ,  0.71400493,  0.36076599, …, -0.36478721,
         0.43214835,  0.1966711 ],
       [-1.73096829,  1.29819543,  0.05857357, …, -0.6101254 ,
        -0.05507101, -0.90948499],
       …,
       [ 1.73096829,  1.29819543,  1.07707397, …,  1.35258008,
        -0.05507101,  0.74974914],
       [ 1.7314013 , -1.38908087, -1.21735002, …, -0.85546358,
         0.06673383, -1.0477545 ],
       [ 1.73183431, -0.80489037,  0.06976588, …,  0.61656553,
        -0.66409522,  0.1966711 ]])
```

[9]:
```
df_standard = pd.DataFrame(df_standard,
             columns=['ID',
                 'Games Played',
    'Minutes Played',
    'Points Per Game',
    'Field Goals Made',
    'Field Goals Attempts',
    'Field Goals Percent',
    '3Points Made',
    '3Points Attempts',
    '3Points Percent',
    'Free Throw Made',
    'Free Throw Attempts',
    'Free Throw Percent',
    'Offensive Rebounds',
    'Defensive Rebounds',
    'Rebounds',
    'Assists',
    'Steals',
    'Blocks',
    'Turnovers'])

df_standard.head()
```

[9]:

	ID	Games Played	Minutes Played	Points Per Game	Field Goals Made	Field Goals Attempts	Field Goals Percent	3Points Made	3Points Attempts	3Points Percent	Free Throw Made	Free Throw Attempts	Free Throw Percent	Offensive Rebounds	Defensive Rebounds	Rebounds	Assists	Steals
0	-1.731834	1.006100	0.640574	0.123403	0.113959	0.047090	0.177269	-0.428374	-0.486911	0.188493	0.655953	0.760387	0.070392	1.428377	-0.121037	0.266040	1.161949	1.107242
1	-1.731401	0.714005	0.360766	0.748626	0.822648	0.465578	1.704465	-1.469853	-1.712288	0.957140	1.087875	1.319370	-0.341888					
2	-1.730968	1.298195	0.058574	-0.640758	-0.535673	-0.482994	-0.293888	0.352736	0.361427	0.919645	-1.071732	-1.076272	0.415557					
3	-1.730535	0.012976	0.058574	0.216029	0.409246	0.130788	1.282049	0.092366	-0.015612	0.257234	-0.531830	-0.357579	-0.428180					
4	-1.730102	0.012976	-0.086926	-0.826009	-0.653788	-0.789885	1.005854	0.613106	0.549947	-0.367682	-1.287692	-1.156127	-1.665021					

```
1            3.210407            1.100107  1.608950 -0.681844 -0.364787
2           -0.608229           -0.264701 -0.405415 -0.608092 -0.610125
3           -0.353653           -0.121037 -0.117649  0.129425 -0.610125
4            1.682952            0.381787  0.793612 -0.903099 -0.610125

     Blocks  Turnovers
0 -0.055071   0.473210
1  0.432148   0.196671
2 -0.055071  -0.909485
3 -0.176876   0.888019
4  0.432148  -0.771215
```

Test Data

[10]:
```
X_test = df_test.copy()
X_test = scaler.fit_transform(X_test)
```

[11]:
```
X_test
```

[11]:
```
array([[-1.73159494, -0.3996569 , -1.09444225, …, -1.10492692,
         0.06609733, -0.64281172],
       [-1.7306831 , -1.15769727,  0.07446345, …, -0.13054963,
        -0.40296321,  0.76098325],
       [-1.72977125,  1.11642385,  1.74760297, …,  1.57461062,
         0.06609733,  1.04174224],
       …,
       [ 1.72977125, -0.57458929, -1.00276337, …, -0.61773827,
        -0.09025619, -1.06395021],
       [ 1.7306831 ,  1.52459944,  2.2518368 , …,  1.57461062,
         0.06609733,  1.60326023],
       [ 1.73159494, -0.4579677 , -0.76210631, …, -0.86133259,
        -0.09025619, -0.08129373]])
```

[12]:
```
X_test = pd.DataFrame(X_test,
             columns=['ID',
                 'Games Played',
    'Minutes Played',
    'Points Per Game',
    'Field Goals Made',
    'Field Goals Attempts',
    'Field Goals Percent',
    '3Points Made',
    '3Points Attempts',
    '3Points Percent',
    'Free Throw Made',
    'Free Throw Attempts',
    'Free Throw Percent',
```

```
    'Offensive Rebounds',
    'Defensive Rebounds',
    'Rebounds',
    'Assists',
    'Steals',
    'Blocks',
    'Turnovers'])

X_test.head()
```

[12]:

	ID	Games Played	Minutes Played	Points Per Game	Field Goals Made	Field Goals Attempts	Field Goals Percent	3Points Made	3Points Attempts	3Points Percent	Free Throw Made	Free Throw Attempts	Free Throw Percent	Offensive Rebounds	Defensive Rebounds	Rebounds	Assists	Steals	Blocks	Turnovers
0	-1.731595	-0.399657	-1.094442	-0.775014	-0.731786	-0.727231	-0.148870	-0.409418	-0.472033	-0.747469	-0.755754	-0.602751	-0.785478	0.132355	-1.005643	-0.761110	-0.925982	-1.104927	0.066097	-0.642812
1	-1.730683	-1.157697	0.074463	0.645520	0.512139	0.502248	0.231965	0.903137	0.857851	0.993635	0.432127	0.437073	0.352629	-0.758712	-0.932743	-0.857711	1.395558	-0.130550	-0.402963	0.760983
2	-1.729771	1.116424	1.747603	0.924970	1.222953	1.200816	0.165733	0.640626	1.047834	1.601142	0.432127	0.597045	-0.039492	0.259650	0.816841	0.591302	0.646674	1.574611	0.066097	1.041742
3	-1.728859	1.349667	2.985268	2.671528	2.348409	2.681780	-0.281334	0.640626	0.952843	-0.359165	3.347837	3.476558	-0.068184	0.514241	0.743942	0.832805	1.844888	0.600233	-0.246610	3.287814
4	-1.727948	-0.283035	-0.727727	-0.612002	-0.731786	-0.643403	-0.761516	0.640626	0.857851	1.219101	-0.323797	-0.522765	0.505652	-1.140598	-1.151442	-1.147514	-0.102210	-0.374144	-1.028377	-0.502432

0.2 Split into training and validation - standardised data

Train data

[13]:
```
# Split randomly the dataset with random_state=8 into 2 different sets:␣
 ↪training data (80%) and validation data (20%)
X_train, X_val, y_train, y_val = train_test_split (df_standard, target_var,␣
 ↪test_size=0.2, random_state=8)

# Save the different sets in the folder `data/processed`
np.save('../data/processed/X_train', X_train)
np.save('../data/processed/X_val',   X_val)
np.save('../data/processed/y_train', y_train)
np.save('../data/processed/y_val',   y_val)
```

[14]:
```
print("Standardised samples:")
print("Dimension of features training data", X_train.shape)
print("Dimension of target training data", y_train.shape)
print("Dimension of features validation data", X_val.shape)
print("Dimension of targer validation data", y_val.shape)
```

```
Standardised samples:
Dimension of features training data (6400, 20)
Dimension of target training data (6400,)
Dimension of features validation data (1600, 20)
Dimension of targer validation data (1600,)
```

[15]:
```
X_train.head()
```

[15]:

	ID	Games Played	Minutes Played	Points Per Game	\
3617	-0.165627	0.071395	-0.523427	-0.687070	
1120	-1.246860	1.064519	0.248843	0.077091	
3873	-0.054776	-0.746471	-1.452389	-1.358606	
153	-1.665583	0.538748	0.080958	-0.524976	
2960	-0.450117	-1.038567	0.125727	-0.270255	

	Field Goals Made	Field Goals Attempts	Field Goals Percent	\
3617	-0.771903	-0.734087	-0.261394	
1120	-0.063214	-0.008708	0.031048	
3873	-1.362477	-1.319970	-1.333681	
153	-0.476616	-0.343498	-0.683810	
2960	-0.181328	-0.092406	-0.407615	

	3Points Made	3Points Attempts	3Points Percent	Free Throw Made	\
3617	-1.209483	-0.863950	-0.480167	0.008072	
1120	0.352736	0.267168	-0.467668	0.439993	
3873	-0.688743	-0.298391	-1.417540	-1.287692	
153	-0.428374	-0.298391	0.357221	-0.531830	
2960	-0.428374	-0.392651	-1.392544	-0.315869	

```
      Free Throw Attempts  Free Throw Percent  Offensive Rebounds  \
3617            -0.038161           -0.207657           -0.608229
1120             0.121549            0.731958            0.791937
3873            -1.395691           -0.265185           -1.244669
153             -0.277725           -2.201944           -0.480941
2960            -0.197870           -0.897987           -0.099078

      Defensive Rebounds  Rebounds   Assists    Steals    Blocks  Turnovers
3617           -0.336533 -0.501337 -0.755595 -0.610125 -0.176876  -0.771215
1120            1.315603  1.129339 -0.681844 -0.119449 -0.176876   0.058402
3873           -1.414013 -1.364637 -0.091830 -0.610125 -0.664095  -1.324294
153            -0.121037 -0.165610 -0.386837 -0.364787 -0.055071  -0.494676
2960           -0.480197 -0.309493 -0.091830 -0.610125  0.066734   0.749749
```

Test

[16]:
```
# Save the different sets in the folder `data/processed`
np.save('../data/processed/X_test',  X_test)
```

[17]:
```
print("Standardised:")
print("Dimension of test data", X_test.shape)
```

```
Standardised:
Dimension of test data (3799, 20)
```

[18]:
```
X_test.head()
```

[18]:
```
         ID  Games Played  Minutes Played  Points Per Game  Field Goals Made  \
0 -1.731595     -0.399657       -1.094442        -0.775014         -0.731786
1 -1.730683     -1.157697        0.074463         0.645520          0.512139
2 -1.729771      1.116424        1.747603         0.924970          1.222953
3 -1.728859      1.349667        2.985268         2.671528          2.348409
4 -1.727948     -0.283035       -0.727727        -0.612002         -0.731786

   Field Goals Attempts  Field Goals Percent  3Points Made  3Points Attempts  \
0             -0.727231            -0.148870     -0.409418         -0.472033
1              0.502248             0.231965      0.903137          0.857851
2              1.200816             0.165733      0.640626          1.047834
3              2.681780            -0.281334      0.640626          0.952843
4             -0.643403            -0.761516      0.640626          0.857851

   3Points Percent  Free Throw Made  Free Throw Attempts  Free Throw Percent  \
0        -0.747469        -0.755754            -0.602751           -0.785478
1         0.993635         0.432127             0.437073            0.352629
2         1.601142         0.432127             0.597045           -0.039492
3        -0.359165         3.347837             3.476558           -0.068184
4         1.219101        -0.323797            -0.522765            0.505652
```

```
   Offensive Rebounds  Defensive Rebounds  Rebounds   Assists    Steals  \
0            0.132355           -1.005643 -0.761110 -0.925982 -1.104927
1           -0.758712           -0.932743 -0.857711  1.395558 -0.130550
2            0.259650            0.816841  0.591302  0.646674  1.574611
3            0.514241            0.743942  0.832805  1.844888  0.600233
4           -1.140598           -1.151442 -1.147514 -0.102210 -0.374144

     Blocks  Turnovers
0  0.066097  -0.642812
1 -0.402963   0.760983
2  0.066097   1.041742
3 -0.246610   3.287814
4 -1.028377  -0.502432
```

0.3 Split into training and validation - non-standardised data

[19]:
```
target_var_o = df_main.pop('TARGET_5Yrs')
target_var_o.head()
```

[19]:
```
0    1
1    1
2    1
3    1
4    1
Name: TARGET_5Yrs, dtype: int64
```

[20]:
```
# Split randomly the dataset with random_state=8 into 2 different sets:␣
 ↪training data (80%) and validation data (20%)
X_train_o, X_val_o, y_train_o, y_val_o = train_test_split (df_main,␣
 ↪target_var_o, test_size=0.2, random_state=8)

# Save the different sets in the folder `data/processed`
np.save('../data/processed/X_train_o', X_train_o)
np.save('../data/processed/X_val_o',   X_val_o)
np.save('../data/processed/y_train_o', y_train_o)
np.save('../data/processed/y_val_o',   y_val_o)
```

[21]:
```
print("Dimension of features training data", X_train_o.shape)
print("Dimension of features validation data", X_val_o.shape)
print("Dimension of target training data", y_train_o.shape)
print("Dimension of targer validation data", y_val_o.shape)
```

```
Dimension of features training data (6400, 20)
Dimension of features validation data (1600, 20)
Dimension of target training data (6400,)
Dimension of targer validation data (1600,)
```

[22]:
```
X_train_o.head()
```

[22]:

	Id	Games Played	Minutes Played	Points Per Game	Field Goals Made	Field Goals Attempts	Field Goals Percent	3Points Made	3Points Attempts	3Points Percent	Free Throw Made	Free Throw Attempts	Free Throw Percent	Offensive Rebounds	Defensive Rebounds	Rebounds	Assists	Steals	Blocks	Turnovers
3617	7416	64	13.9	4.3	1.5	3.6	43.0	-0.2	-0.1	11.9	1.4	1.9	69.2	0.6	1.7	2.2	0.6	0.4	0.1	0.7
1120	4919	81	20.8	7.6	2.7	6.2	44.8	0.4	1.1	12.1	1.8	2.1	79.0	1.7	4.0	5.6	0.7	0.6	0.1	1.3
3873	7672	50	5.6	1.4	0.5	1.5	36.4	0.0	0.5	-3.1	0.2	0.2	68.6	0.1	0.2	0.4	1.5	0.4	-0.3	0.3
153	3952	72	19.3	5.0	2.0	5.0	40.4	0.1	0.5	25.3	0.9	1.6	48.4	0.7	2.0	2.9	1.1	0.5	0.2	0.9
2960	6759	45	19.7	6.1	2.5	5.9	42.1	0.1	0.4	-2.7	1.1	1.7	62.0	1.0	1.5	2.6	1.5	0.4	0.3	1.8

[ ]:
```

```