МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Лабораторная работа №3
по курсу «Методы машинного обучения»

Тема: «Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных»

ИСПОЛНИТЕЛЬ: ___Сметанкин К.И__
ФИО
группа ИУ5-22М _____
подпись
"__"_____2020 г.

ПРЕПОДАВАТЕЛЬ: ___Гапанюк Ю.Е___
ФИО
_____
подпись
"__"_____2020 г.

Москва - 2020

In [0]:

```
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer

%matplotlib inline
sns.set(style="ticks")
```

In [0]:

```
url = 'https://raw.githubusercontent.com/Smet1/bmstu_ml/master/lab3/data.csv'
df = pd.read_csv(url, error_bad_lines=False)
```

In [32]:

```
df.head()
```

Out[32]:

	Unnamed: 0	ID	Name	Age	Photo	Nationality	
0	0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	h
1	1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	h
2	2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	h
3	3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	h
4	4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	

5 rows × 89 columns

In [33]:

```
row_number = df.shape[0]
column_number = df.shape[1]

print('Данный датасет содержит {} строк и {} столбца.'.format(row_number, column_numbe
r))
```

Данный датасет содержит 18207 строк и 89 столбца.

1. Обработка пропусков в данных

In [34]:

```
for col in df.columns:
    null_count = df[df[col].isnull()].shape[0]
    if null_count > 0:
        column_type = df[col].dtype
        percent = round((null_count / row_number) * 100, 3)
        print('{} - {} - {}. Тип - {}'.format(col, null_count, percent, column_ty
pe))
```

Club - 241 - 1.324. Тип - object
Preferred Foot - 48 - 0.264. Тип - object
International Reputation - 48 - 0.264. Тип - float64
Weak Foot - 48 - 0.264. Тип - float64
Skill Moves - 48 - 0.264. Тип - float64
Work Rate - 48 - 0.264. Тип - object
Body Type - 48 - 0.264. Тип - object
Real Face - 48 - 0.264. Тип - object
Position - 60 - 0.33. Тип - object
Jersey Number - 60 - 0.33. Тип - float64
Joined - 1553 - 8.53. Тип - object
Loaned From - 16943 - 93.058. Тип - object
Contract Valid Until - 289 - 1.587. Тип - object
Height - 48 - 0.264. Тип - object
Weight - 48 - 0.264. Тип - object
LS - 2085 - 11.452. Тип - object
ST - 2085 - 11.452. Тип - object
RS - 2085 - 11.452. Тип - object
LW - 2085 - 11.452. Тип - object
LF - 2085 - 11.452. Тип - object
CF - 2085 - 11.452. Тип - object
RF - 2085 - 11.452. Тип - object
RW - 2085 - 11.452. Тип - object
LAM - 2085 - 11.452. Тип - object
CAM - 2085 - 11.452. Тип - object
RAM - 2085 - 11.452. Тип - object
LM - 2085 - 11.452. Тип - object
LCM - 2085 - 11.452. Тип - object
CM - 2085 - 11.452. Тип - object
RCM - 2085 - 11.452. Тип - object
RM - 2085 - 11.452. Тип - object
LWB - 2085 - 11.452. Тип - object
LDM - 2085 - 11.452. Тип - object
CDM - 2085 - 11.452. Тип - object
RDM - 2085 - 11.452. Тип - object
RWB - 2085 - 11.452. Тип - object
LB - 2085 - 11.452. Тип - object
LCB - 2085 - 11.452. Тип - object
CB - 2085 - 11.452. Тип - object
RCB - 2085 - 11.452. Тип - object
RB - 2085 - 11.452. Тип - object
Crossing - 48 - 0.264. Тип - float64
Finishing - 48 - 0.264. Тип - float64
HeadingAccuracy - 48 - 0.264. Тип - float64
ShortPassing - 48 - 0.264. Тип - float64
Volleys - 48 - 0.264. Тип - float64
Dribbling - 48 - 0.264. Тип - float64
Curve - 48 - 0.264. Тип - float64
FKAccuracy - 48 - 0.264. Тип - float64
LongPassing - 48 - 0.264. Тип - float64
BallControl - 48 - 0.264. Тип - float64
Acceleration - 48 - 0.264. Тип - float64
SprintSpeed - 48 - 0.264. Тип - float64
Agility - 48 - 0.264. Тип - float64
Reactions - 48 - 0.264. Тип - float64
Balance - 48 - 0.264. Тип - float64
ShotPower - 48 - 0.264. Тип - float64
Jumping - 48 - 0.264. Тип - float64
Stamina - 48 - 0.264. Тип - float64
Strength - 48 - 0.264. Тип - float64
LongShots - 48 - 0.264. Тип - float64

```
Aggression - 48 - 0.264. Тип - float64
Interceptions - 48 - 0.264. Тип - float64
Positioning - 48 - 0.264. Тип - float64
Vision - 48 - 0.264. Тип - float64
Penalties - 48 - 0.264. Тип - float64
Composure - 48 - 0.264. Тип - float64
Marking - 48 - 0.264. Тип - float64
StandingTackle - 48 - 0.264. Тип - float64
SlidingTackle - 48 - 0.264. Тип - float64
GKDiving - 48 - 0.264. Тип - float64
GKHandling - 48 - 0.264. Тип - float64
GKKicking - 48 - 0.264. Тип - float64
GKPositioning - 48 - 0.264. Тип - float64
GKReflexes - 48 - 0.264. Тип - float64
Release Clause - 1564 - 8.59. Тип - object
```

1.1 Удаление пустых значений

In [0]:

```
df = df[df['Club'].notna()]

# удаление столбца
df.drop(columns=['Loaned From'], inplace=True)
```

In [36]:

```
row_number = df.shape[0]
column_number = df.shape[1]

print('Данный датасет содержит {} строк и {} столбца.'.format(row_number, column_numbe
r))
```

```
Данный датасет содержит 17966 строк и 88 столбца.
```

1.2 Заполнение нулями

In [37]:

```
df['FKAccuracy'] = df['FKAccuracy'].fillna(0)

df[df['FKAccuracy'].isnull()].shape
```

Out[37]:

```
(0, 88)
```

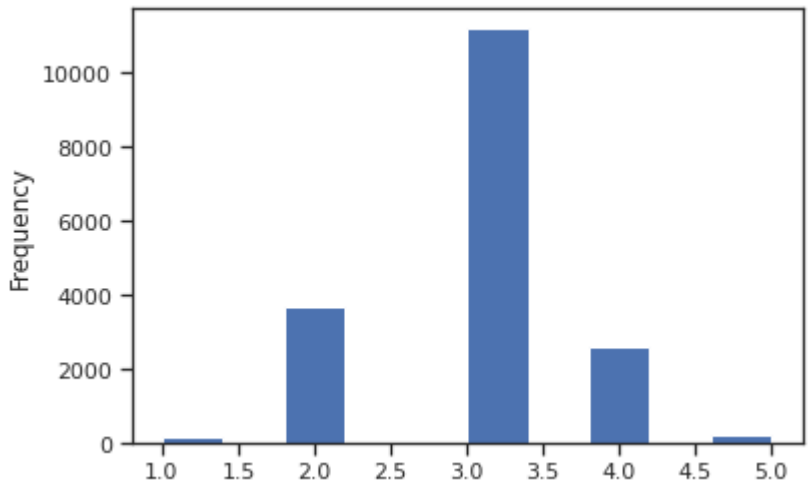
1.3 Внедрение значений в числовых данных

In [44]:

```
df['Weak Foot'].plot.hist()
```

Out[44]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7719ae2ba8>
```



In [45]:

```
df['Weak Foot'].describe()
```

Out[45]:

```
count    17918.000000
mean         2.947260
std          0.660106
min          1.000000
25%          3.000000
50%          3.000000
75%          3.000000
max          5.000000
Name: Weak Foot, dtype: float64
```

In [46]:

```
mode = df['Weak Foot'].mode()[0]
(df[df['Weak Foot'] == mode].shape[0]/row_number) * 100
```

Out[46]:

```
62.34554157853724
```

In [47]:

```
median = df['Weak Foot'].describe()['50%']
(df[df['Weak Foot'] == median].shape[0]/row_number) * 100
```

Out[47]:

```
62.34554157853724
```

In [0]:

```
imp = SimpleImputer(strategy='most_frequent')
df['Weak Foot'] = imp.fit_transform(df[['Weak Foot']])
```

In [50]:

```
df[df['Weak Foot'].isnull()].shape
```

Out[50]:

```
(0, 88)
```

1.4 Внедрение значений в категориальных данных

In [0]:

```
imp = SimpleImputer(strategy='most_frequent')
df['Stamina'] = imp.fit_transform(df[['Stamina']])
```

In [0]:

```

```

In [43]:

```
df[df['Stamina'].isnull()].shape
```

Out[43]:

```
(0, 88)
```

2. Кодирование категориальных признаков

In [56]:

```
for col in df.columns:
    column_type = df[col].dtype
    if column_type == 'object':
        print(col)
```

```
Name
Photo
Nationality
Flag
Club
Club Logo
Value
Wage
Preferred Foot
Work Rate
Body Type
Real Face
Position
Joined
Contract Valid Until
Height
Weight
LS
ST
RS
LW
LF
CF
RF
RW
LAM
CAM
RAM
LM
LCM
CM
RCM
RM
LWB
LDM
CDM
RDM
RWB
LB
LCB
CB
RCB
RB
Release Clause
```

2.1 Кодирование категорий целочисленными значениями

In [64]:

```
df['Club'].unique()
```

Out[64]:

```
array(['FC Barcelona', 'Juventus', 'Paris Saint-Germain',
       'Manchester United', 'Manchester City', 'Chelsea', 'Real Madr
id',
       'Atlético Madrid', 'FC Bayern München', 'Tottenham Hotspur',
       'Liverpool', 'Napoli', 'Arsenal', 'Milan', 'Inter', 'Lazio',
       'Borussia Dortmund', 'Vissel Kobe', 'Olympique Lyonnais', 'Ro
ma',
       'Valencia CF', 'Guangzhou Evergrande Taobao FC', 'FC Porto',
       'FC Schalke 04', 'Beşiktaş JK', 'LA Galaxy', 'Sporting CP',
       'Real Betis', 'Olympique de Marseille', 'RC Celta',
       'Bayer 04 Leverkusen', 'Real Sociedad', 'Villarreal CF',
       'Sevilla FC', 'SL Benfica', 'AS Saint-Étienne', 'AS Monaco',
       'Leicester City', 'Atalanta', 'Grêmio', 'Atlético Mineiro',
       'RB Leipzig', 'Ajax', 'Dalian YiFang FC', 'Everton',
       'West Ham United', '1. FC Köln', 'TSG 1899 Hoffenheim',
       'Shanghai SIPG FC', 'OGC Nice', 'Al Nassr',
       'Wolverhampton Wanderers', 'Borussia Mönchengladbach',
       'Hertha BSC', 'SV Werder Bremen', 'Cruzeiro',
       'Athletic Club de Bilbao', 'Torino', 'Medipol Başakşehir FK',
       'Beijing Sinobo Guoan FC', 'Crystal Palace', 'PFC CSKA Mosco
w',
       'VfL Wolfsburg', 'Shakhtar Donetsk', 'Toronto FC',
       'Lokomotiv Moscow', 'Sassuolo', 'New York City FC', 'Fluminen
se',
       'PSV', 'Levante UD', 'Fulham', 'Watford', 'Atlanta United',
       'Montpellier HSC', 'Galatasaray SK', 'Fenerbahçe SK', 'SD Eib
ar',
       'Los Angeles FC', 'Sampdoria', 'Al Hilal', 'VfB Stuttgart',
       'SC Braga', 'River Plate', 'Deportivo Alavés',
       'Eintracht Frankfurt', 'Girona FC', 'Guangzhou R&F; FC', 'Bur
nley',
       'Stoke City', 'Southampton', 'Tianjin Quanjian FC', 'Getafe C
F',
       'Beijing Renhe FC', 'Montreal Impact', 'Chievo Verona', 'Geno
a',
       'Portland Timbers', 'Tigres U.A.N.L.', 'RCD Espanyol',
       'Hebei China Fortune FC', 'Cagliari', 'Chicago Fire', 'DC Uni
ted',
       'Sagan Tosu', 'Dynamo Kyiv', 'Santos', 'Internacional',
       'América FC (Minas Gerais)', 'Independiente', 'Boca Juniors',
       'Cruz Azul', '1. FSV Mainz 05', 'Bournemouth', 'Spartak Mosco
w',
       'Racing Club', 'FC Augsburg', 'Fiorentina', 'FC Nantes',
       'Feyenoord', 'Club Brugge KV', 'Brighton & Hove Albion', 'Al
Ahli',
       'Jiangsu Suning FC', 'SC Freiburg', 'PAOK', 'Stade Rennais F
C',
       'Trabzonspor', 'SPAL', 'Portimonense SC', 'Olympiacos CFP',
       'Club Atlético Huracán', 'Kasimpaşa SK', 'Newcastle United',
       'Frosinone', 'Querétaro', 'KRC Genk', 'Hannover 96',
       'Stade Malherbe Caen', 'Godoy Cruz', 'Toulouse Football Clu
b',
       'RSC Anderlecht', 'Huddersfield Town', 'CD Tondela',
       'Seattle Sounders FC', 'Hamburger SV', 'FC Red Bull Salzbur
g',
       'Rio Ave FC', 'FC Girondins de Bordeaux', 'Melbourne Victor
y',
       'Parma', 'FC Basel 1893', 'Al Wehda', 'BSC Young Boys', 'KAA
Gent',
```

```
       'Al Ittihad', 'Standard de Liège', 'Shanghai Greenland Shenhu
a FC',
       'Colo-Colo', 'Junior FC', 'West Bromwich Albion',
       'RC Strasbourg Alsace', 'Göztepe SK', 'Deportivo Cali',
       'Deportivo Toluca', 'Bologna', 'Nagoya Grampus', 'Amiens SC',
       'Changchun Yatai FC', 'Club Atlético Lanús', 'Botafogo',
       'Club América', 'Udinese', 'Real Valladolid CF', 'CD Legané
s',
       'Club Atlético Banfield', 'Celtic', 'Vitória Guimarães',
       'FC København', 'UD Las Palmas', 'Deportivo de La Coruña',
       'Universidad Católica', 'San Lorenzo de Almagro', 'Rayo Valle
cano',
       'Monterrey', 'Columbus Crew SC', 'MKE Ankaragücü',
       'Guizhou Hengfeng FC', 'Swansea City', 'Tianjin TEDA FC',
       'Chongqing Dangdai Lifan FC SWM Team', 'AEK Athens', 'Al Taaw
oun',
       'Melbourne City FC', 'En Avant de Guingamp',
       'Akhisar Belediyespor', 'Foggia', 'LOSC Lille', '1. FC Nürnbe
rg',
       'Clube Sport Marítimo', 'Real Sporting de Gijón', 'BB Erzurum
spor',
       'Shandong Luneng TaiShan FC', 'Club Atlético Colón', 'Bahia',
       'Once Caldas', 'FC Groningen', 'Angers SCO', 'Paraná',
       'Antalyaspor', 'Minnesota United FC', 'Club León', 'Empoli',
       'VVV-Venlo', 'Leeds United', 'Viktoria Plzeň', 'Alanyaspor',
       'Atlético Paranaense', 'Derby County', 'Kawasaki Frontale',
       'Cardiff City', 'Aston Villa', 'Guadalajara', 'Dijon FCO',
       'Santos Laguna', 'Málaga CF', 'Vitória', 'Çaykur Rizespor',
       'U.N.A.M.', 'Nottingham Forest', 'Royal Antwerp FC',
       'Club Tijuana', 'Sport Club do Recife', 'Real Salt Lake',
       'AZ Alkmaar', 'SK Slavia Praha', 'Willem II', 'Middlesbroug
h',
       'Dinamo Zagreb', 'Club Atlas', 'Granada CF', 'Sydney FC',
       'Sporting Kansas City', 'SV Zulte-Waregem', 'Philadelphia Uni
on',
       'Real Oviedo', 'Pachuca', 'Boavista FC', 'Atiker Konyaspor',
       'Kaizer Chiefs', 'GD Chaves', 'Palermo', 'Atlético Nacional',
       'Puebla FC', 'Perth Glory', 'Panathinaikos FC', 'FC Sion',
       'Vitória de Setúbal', 'New York Red Bulls', 'Al Shabab',
       'Monarcas Morelia', 'Albacete BP', 'Rangers FC', 'Sparta Prah
a',
       'Legia Warszawa', 'Urawa Red Diamonds', 'Rosario Central',
       'Stade de Reims', 'ADO Den Haag', 'Chapecoense', 'FC Midtjyll
and',
       'San Jose Earthquakes', 'Belgrano de Córdoba', 'Brescia',
       'Kashima Antlers', 'CD Everton de Viña del Mar',
       'Fortuna Düsseldorf', 'SD Huesca', 'Preston North End',
       'Club Atlético Talleres', 'Benevento', 'Vitesse',
       'Gimnasia y Esgrima La Plata', 'Houston Dynamo', 'Club Necax
a',
       'Norwich City', 'Holstein Kiel', 'Ettifaq FC', 'Kayserispor',
       '1. FC Heidenheim 1846', 'Brentford', 'Yeni Malatyaspor',
       'Lobos BUAP', 'Bursaspor', 'Ceará Sporting Club',
       'Sheffield United', 'FC Ingolstadt 04', 'Estudiantes de La Pl
ata',
       'AIK', 'Queens Park Rangers', 'Suwon Samsung Bluewings',
       'Heart of Midlothian', 'Reading', 'FC Dallas', 'Heracles Alme
lo',
       'Venezia FC', 'CD Lugo', 'Henan Jianye FC', 'Orlando City S
C',
       'CA Osasuna', 'NAC Breda', 'Livorno', 'Universidad de Chile',
```

```
       'Brøndby IF', 'Aberdeen', 'Defensa y Justicia', 'Atlético Tuc
umán',
       'Blackburn Rovers', 'SV Darmstadt 98', 'Moreirense FC',
       'Sanfrecce Hiroshima', 'CD Numancia', 'KV Oostende', 'FC Utre
cht',
       'Vancouver Whitecaps FC', 'Odense Boldklub', 'SC Heerenveen',
       'Racing Club de Lens', 'Independiente Santa Fe',
       'Sporting de Charleroi', 'Millonarios FC', 'Sheffield Wednesd
ay',
       'Perugia', 'Daegu FC', 'Vélez Sarsfield',
       'Grasshopper Club Zürich', 'Sivasspor', 'Nîmes Olympique',
       'Rosenborg BK', 'SK Sturm Graz', 'FC Metz',
       'CD Universidad de Concepción', 'Hellas Verona', 'Brisbane Ro
ar',
       'CD Feirense', 'Hull City', 'Waasland-Beveren', 'Neuchâtel Xa
max',
       'Real Zaragoza', 'CD Aves', 'Millwall', 'Unión de Santa Fe',
       'KAS Eupen', 'Cádiz CF', 'FC Tokyo', 'CD Tenerife',
       '1. FC Union Berlin', 'Al Fayha', 'AJ Auxerre',
       'Patriotas Boyacá FC', 'Molde FK', 'Bristol City', 'CD Nacion
al',
       'Sporting Lokeren', 'FC St. Pauli', 'Deportes Iquique',
       'Al Qadisiyah', 'Atlético Bucaramanga', 'Club Atlético Tigr
e',
       'FK Austria Wien', 'Patronato', 'Malmö FF', 'Kashiwa Reysol',
       'US Cremonese', 'VfL Bochum 1848', 'SK Rapid Wien',
       'KSV Cercle Brugge', 'Rionegro Águilas', 'Gimnàstic de Tarrag
ona',
       'Lecce', 'Santa Clara', 'BK Häcken', 'New England Revolutio
n',
       'Orlando Pirates', 'Atlético Huila', 'Western Sydney Wanderer
s',
       'Kalmar FF', 'Independiente Medellín', 'Fortuna Sittard',
       'Lech Poznań', 'Djurgårdens IF', 'CF Reus Deportiu', 'SK Bran
n',
       'Ulsan Hyundai FC', 'Sint-Truidense VV', 'Carpi', 'Al Fateh',
       'Royal Excel Mouscron', 'AC Ajaccio', 'PEC Zwolle', 'Sunderla
nd',
       'Club Atlético Aldosivi', 'US Salernitana 1919', 'FC Lorien
t',
       'Argentinos Juniors', 'AD Alcorcón', 'Crotone', 'Excelsior',
       'KV Kortrijk', 'IFK Norrköping', 'Adelaide United',
       'FC St. Gallen', 'Tiburones Rojos de Veracruz', 'CD Palestin
o',
       'Jeju United FC', 'Deportes Tolima', 'Jeonbuk Hyundai Motor
s',
       'Birmingham City', 'América de Cali', 'La Equidad', 'Spezia',
       'Aalborg BK', 'Le Havre AC', 'Górnik Zabrze',
       'Central Coast Mariners', 'Wigan Athletic',
       'Jagiellonia Białystok', 'Cittadella', 'Hibernian', 'FC Lugan
o',
       'San Martín de San Juan', 'Strømsgodset IF', 'Júbilo Iwata',
       "Newell's Old Boys", 'Al Faisaly', 'Colorado Rapids',
       'IF Elfsborg', 'SV Sandhausen', 'Al Batin', 'Stade Brestois 2
9',
       'UD Almería', 'Gyeongnam FC', 'Yokohama F. Marinos', 'Kilmarn
ock',
       'Pescara', 'Newcastle Jets', 'Córdoba CF', 'RCD Mallorca',
       'Hammarby IF', 'Cerezo Osaka', 'KFC Uerdingen 05',
       'Shimizu S-Pulse', 'MSV Duisburg', 'Os Belenenses',
       'DSC Arminia Bielefeld', 'Ipswich Town', 'FC Seoul',
```

```
       'Lechia Gdańsk', 'Gamba Osaka', 'CF Rayo Majadahonda', 'LASK
Linz',
       'Bolton Wanderers', 'Al Raed', 'Extremadura UD', 'SC Paderbor
n 07',
       'Wellington Phoenix', 'Unión Española', 'Alianza Petrolera',
       'Cracovia', 'Gangwon FC', 'Elche CF', 'ESTAC Troyes', 'AS Béz
iers',
       'La Berrichonne de Châteauroux', 'Clermont Foot 63',
       '1. FC Magdeburg', 'Pohang Steelers', 'Örebro SK', 'Arka Gdyn
ia',
       'SG Dynamo Dresden', 'SpVgg Greuther Fürth', 'CD Huachipato',
       'Wisła Kraków', 'Stabæk Fotball', 'Eintracht Braunschweig',
       'Valenciennes FC', 'FC Thun', 'San Luis de Quillota',
       ' SSV Jahn Regensburg', 'Cosenza', 'FC Nordsjælland',
       'FC Erzgebirge Aue', 'Jeonnam Dragons', 'Wolfsberger AC',
       'Chamois Niortais Football Club', 'Club Deportes Temuco',
       'AS Nancy Lorraine', 'Red Star FC', 'Al Hazem', 'Pogoń Szczec
in',
       'Charlton Athletic', 'Grenoble Foot 38', 'FC Hansa Rostock',
       'San Martin de Tucumán', 'Incheon United FC', 'Śląsk Wrocła
w',
       'GFC Ajaccio', '1. FC Kaiserslautern', 'Deportivo Pasto',
       'Lincoln City', 'Motherwell', 'Rotherham United', 'Burton Alb
ion',
       'Wisła Płock', 'FC Wacker Innsbruck', 'Peterborough United',
       'Ascoli', 'FC Zürich', 'Fleetwood Town', 'Padova',
       'FC Sochaux-Montbéliard', 'SV Wehen Wiesbaden', 'Unión La Cal
era',
       'Scunthorpe United', "CD O'Higgins", 'CD Antofagasta',
       'Plymouth Argyle', 'Aarhus GF', 'Lillestrøm SK', 'Karlsruher
SC',
       'GIF Sundsvall', 'FC Emmen', 'Barnsley', 'Audax Italiano',
       'V-Varen Nagasaki', 'Paris FC', 'SpVgg Unterhaching', 'Hobro
IK',
       'De Graafschap', 'Hokkaido Consadole Sapporo', 'Tromsø IL',
       'FC Luzern', 'FK Haugesund', 'Zagłębie Lubin', 'VfR Aalen',
       'Dundalk', 'Oxford United', 'Piast Gliwice', 'Ohod Club',
       'Östersunds FK', 'Vegalta Sendai', 'Crawley Town',
       'FC Admira Wacker Mödling', 'Vålerenga Fotball', 'Dundee FC',
       'Portsmouth', 'Envigado FC', 'Miedź Legnica', 'Odds BK',
       'SC Fortuna Köln', 'US Orléans Loiret Football', 'Sarpsborg 0
8 FF',
       'Jaguares de Córdoba', 'Bradford City', 'Accrington Stanley',
       'St. Johnstone FC', 'Boyacá Chicó FC', 'Luton Town',
       'SV Mattersburg', 'Kristiansund BK', 'Sangju Sangmu FC',
       'Rochdale', 'Walsall', 'Korona Kielce', 'Shonan Bellmare',
       'FC Würzburger Kickers', 'FSV Zwickau', 'St. Mirren', 'AC Hor
sens',
       'Esbjerg fB', 'HJK Helsinki', 'Southend United', 'Bristol Rov
ers',
       'Hamilton Academical FC', 'TSV 1860 München', 'Curicó Unido',
       'SCR Altach', 'Ranheim Fotball', 'Stevenage',
       'SG Sonnenhof Großaspach', 'Oldham Athletic', 'Milton Keynes
Dons',
       'FK Bodø/Glimt', 'SC Preußen Münster', 'Wycombe Wanderers',
       'Vejle Boldklub', 'Bury', 'Randers FC', 'VfL Osnabrück',
       'SønderjyskE', 'IFK Göteborg', 'Mansfield Town', 'Coventry Ci
ty',
       'Waterford FC', 'Shrewsbury', 'IK Start', 'Gillingham',
       'FC Energie Cottbus', 'FC Carl Zeiss Jena', 'Hallescher FC',
       'SV Meppen', 'AFC Wimbledon', 'Blackpool', 'Doncaster Rover
```

```
s',
       'Sandefjord Fotball', 'VfL Sportfreunde Lotte', 'Cheltenham T
own',
       'IK Sirius', 'Vendsyssel FF', 'Swindon Town', 'Notts County',
       'SKN St. Pölten', 'Exeter City', 'Northampton Town',
       'Shamrock Rovers', 'Colchester United', 'Livingston FC',
       'TSV Hartberg', 'Tranmere Rovers', 'Cambridge United',
       'Grimsby Town', 'Port Vale', 'Itagüí Leones FC',
       'Forest Green Rovers', 'Dalkurd FF', 'Zagłębie Sosnowiec',
       'Carlisle United', 'Trelleborgs FF', "St. Patrick's Athleti
c",
       'Morecambe', 'Cork City', 'IF Brommapojkarna', 'Crewe Alexand
ra',
       'Yeovil Town', 'Bohemian FC', 'Macclesfield Town',
       'Newport County', 'Sligo Rovers', 'Derry City', 'Limerick F
C',
       'Bray Wanderers'], dtype=object)
```

In [0]:

```
le = LabelEncoder()
df['Club_LabelEncoder'] = le.fit_transform(df['Club'])
```

In [79]:

```
df['Club_LabelEncoder'].unique()
```

Out[79]:

```
array([212, 326, 435, 375, 374, 134, 470,  61, 214, 583, 363, 398,
52,
       382, 315, 351,  86, 620, 418, 482, 605, 280, 232, 234,  77, 3
46,
       552, 469, 419, 457,  72, 473, 619, 530, 504,  19,  17, 358,
55,
       278,  62, 456,  26, 176, 206, 633,   3, 574, 535, 412,  36, 6
40,
        87, 297, 511, 168,  56, 581, 377,  74, 169, 427, 616, 531, 5
82,
       367, 527, 401, 254, 428, 359, 260, 630,  58, 390, 264, 250, 4
95,
       368, 514,  34, 612, 488, 480, 182, 198, 272, 281, 100, 566, 5
44,
       577, 268,  73, 391, 137, 267, 450, 580, 459, 293, 121, 136, 1
73,
       513, 195, 524, 316,  46, 312,  82, 167,   7,  89, 549, 462, 2
11,
       252, 230, 251, 150,  95,  28, 324, 490, 425, 562, 585, 505, 4
49,
       417, 146, 340, 404, 259, 455, 331, 291, 561, 273, 584, 461, 3
03,
       117, 529, 288, 233, 478, 220, 379, 436, 213,  41,  69, 328,
35,
       564, 534, 157, 325, 632, 458, 285, 183, 185,  84, 397,  45, 1
31,
       147,  88, 141, 595, 475, 110, 144, 127, 623, 224, 591, 186, 5
97,
       516, 467, 389, 159, 370, 282, 570, 578, 138,  12,  40, 378, 2
01,
        27, 255, 348,   5, 155, 474,  67, 533, 145,  70, 420, 221,
48,
       433,  49, 386, 152, 200, 604, 356, 618,  42,  64, 187, 341, 1
23,
        54, 279, 189, 525, 395, 622, 647, 589, 409, 486, 154, 551, 4
72,
        20, 501, 636, 380, 190, 142, 274, 572, 553, 512, 443, 471, 4
29,
        81,  57, 335, 261, 431,  63, 453, 439, 432, 236, 624, 402,
39,
       388,  43, 465, 548, 357, 602, 483, 563,  11, 132, 229, 515,
75,
        94, 338, 107, 257, 496, 452, 148,  76, 621, 270, 302, 153, 4
08,
       301, 205, 342,   1,  93, 642, 366, 101, 126, 536, 223, 204,
14,
       454, 569, 292, 468, 216, 296, 611, 111, 295, 421, 104, 396, 3
65,
       598,  99,  23, 179,  65,  79, 506, 393, 521, 113, 334, 242, 6
07,
       414, 491, 463, 314, 555, 383, 537, 440, 175, 626, 275, 542, 4
11,
       484, 502, 228, 118, 294,  96, 108, 304, 627, 399, 476, 106, 3
84,
       601, 329, 171, 241, 116,   6,  32,  15, 437, 387,  97, 112, 5
54,
       239, 180,  37,  59, 149, 246, 438, 373, 339, 592, 613, 500, 3
32,
       479, 271, 353, 523,  68, 400, 422,  60, 634, 336, 313, 258, 3
```

```
54,
       191, 120, 499, 596, 541, 125,  31, 487,   8, 426, 568, 143, 5
94,
       225,  50,  10, 166, 207, 333, 308,  25, 238, 579, 115, 321, 1
81,
       322,  78,  47, 350, 550,  21, 352, 284, 128, 635, 319, 139, 2
98,
       226, 519, 567, 327, 405,  30, 158, 306, 509,  29, 560, 590, 2
83,
       644, 343, 441, 403, 172, 460, 290, 129, 330, 538, 371, 423, 1
74,
       317, 235, 355, 265, 119, 347,  85,  38, 209, 492, 631, 599,
44,
       163, 266, 199, 196,  16, 349, 140,   4, 447, 648,  51, 497, 5
46,
       109, 637, 559, 197, 606, 240, 517,   0, 161, 231, 219, 323, 6
39,
       130, 151,  18, 477,  33, 446, 133, 276, 222, 518, 311, 650, 2
62,
         2, 184, 362, 394, 485, 102, 638, 243, 442,  53, 245, 253, 4
30,
       237, 510, 600, 528, 114, 105, 445,  22, 360, 337, 263, 217,
71,
        66, 603, 434, 547, 299, 178, 300, 588, 227, 248, 645, 617, 1
93,
       424, 444, 415, 649, 608, 164, 210, 625, 194, 451, 202, 381, 4
13,
       489, 593, 526, 320,  91,  24, 556,  90, 369, 507, 345, 522, 4
81,
       628, 344, 539, 244, 249, 557,   9, 203, 286, 545,  98, 289, 5
75,
       170, 494, 466, 565, 498, 416, 385, 247, 493, 641, 609, 103, 4
64,
       614, 573, 307, 376, 162, 629, 540, 310, 269, 218, 215, 287, 5
08,
        13,  80, 192, 520, 615, 135, 309, 610, 571, 410, 503, 208, 4
07,
       532, 156, 364, 576, 586, 122, 277, 448, 318, 256, 177, 646, 1
24,
       587, 558, 392, 160, 305, 165, 643,  83, 372, 406, 543, 188, 3
61,
        92])
```

In [81]:

```
le.inverse_transform([212, 305])
```

Out[81]:

```
array(['FC Barcelona', 'IF Brommapojkarna'], dtype=object)
```

In [83]:

```
df[['Club','Club_LabelEncoder']].head()
```

Out[83]:

	Club	Club_LabelEncoder
0	FC Barcelona	212
1	Juventus	326
2	Paris Saint-Germain	435
3	Manchester United	375
4	Manchester City	374

2.2. Кодирование категорий наборами бинарных значений

In [94]:

```
df['Position'].unique()
```

Out[94]:

```
array(['RF', 'ST', 'LW', 'GK', 'RCM', 'LF', 'RS', 'RCB', 'LCM', 'C
B',
       'LDM', 'CAM', 'CDM', 'LS', 'LCB', 'RM', 'LAM', 'LM', 'LB', 'R
DM',
       'RW', 'CM', 'RB', 'RAM', 'CF', 'RWB', 'LWB', nan], dtype=obje
ct)
```

In [99]:

```
# Импьютация наиболее частыми значениями
imp = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
df['Position'] = imp2.fit_transform(df[['Position']])

df['Position']
```

Out[99]:

```
0          RF
1          ST
2          LW
3          GK
4         RCM
         ...
18202      CM
18203      ST
18204      ST
18205      RW
18206      CM
Name: Position, Length: 17966, dtype: object
```

In [100]:

```
# Пустые значения отсутствуют
np.unique(df['Position'])
```

Out[100]:

```
array(['CAM', 'CB', 'CDM', 'CF', 'CM', 'GK', 'LAM', 'LB', 'LCB', 'LC
M',
       'LDM', 'LF', 'LM', 'LS', 'LW', 'LWB', 'RAM', 'RB', 'RCB', 'RC
M',
       'RDM', 'RF', 'RM', 'RS', 'RW', 'RWB', 'ST'], dtype=object)
```

In [0]:

```
ohe = OneHotEncoder()
transformed_data = ohe.fit_transform(df[['Position']])
```

In [102]:

```
transformed_data.shape
```

Out[102]:

```
(17966, 27)
```

In [103]:

```
transformed_data.todense()[0:10]
```

Out[103]:

```
matrix([[0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
0.,
         0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
0.,
         0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1.],
        [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0.,
0.,
         0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
0.,
         0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
0.,
         0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0.,
0.,
         0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
0.,
         0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
0.,
         0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0.],
        [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
0.,
         0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0.],
        [0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
0.,
         0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.]])
```

2.3 Pandas get_dummies

In [104]:

```
pd.get_dummies(df['Position']).head()
```

Out[104]:

	CAM	CB	CDM	CF	CM	GK	LAM	LB	LCB	LCM	LDM	LF	LM	LS	LW	LWB	RAM
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

In [105]:

```
pd.get_dummies(df['Position'], dummy_na=True).head()
```

Out[105]:

	CAM	CB	CDM	CF	CM	GK	LAM	LB	LCB	LCM	LDM	LF	LM	LS	LW	LWB	RAM
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

3. Масштабирование данных

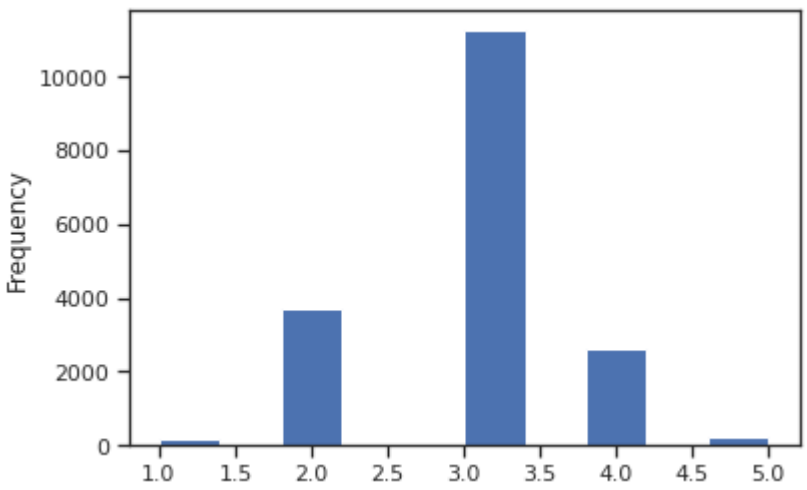
3.1 MinMax масштабирование

In [106]:

```
df['Weak Foot'].plot.hist()
```

Out[106]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7719a109b0>
```



In [0]:

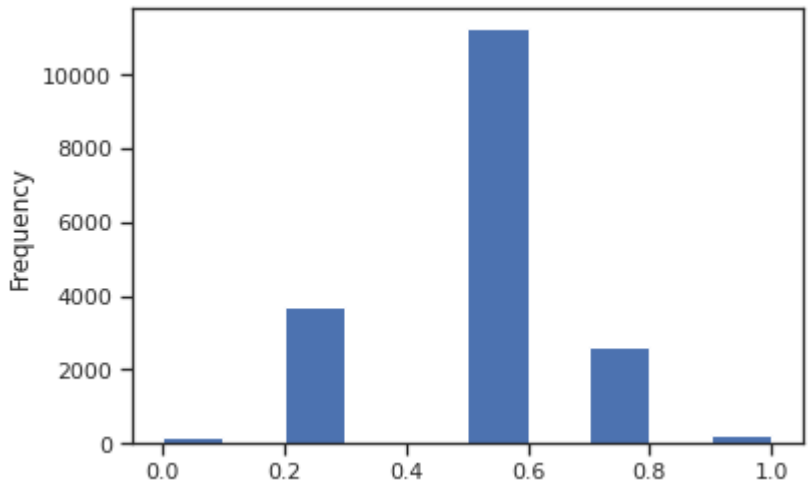
```
sc1 = MinMaxScaler()
df['Weak Foot'] = sc1.fit_transform(df[['Weak Foot']])
```

In [109]:

```
df['Weak Foot'].plot.hist()
```

Out[109]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f77182ddef0>
```



3.2 Масштабирование данных на основе Z-оценки

In [0]:

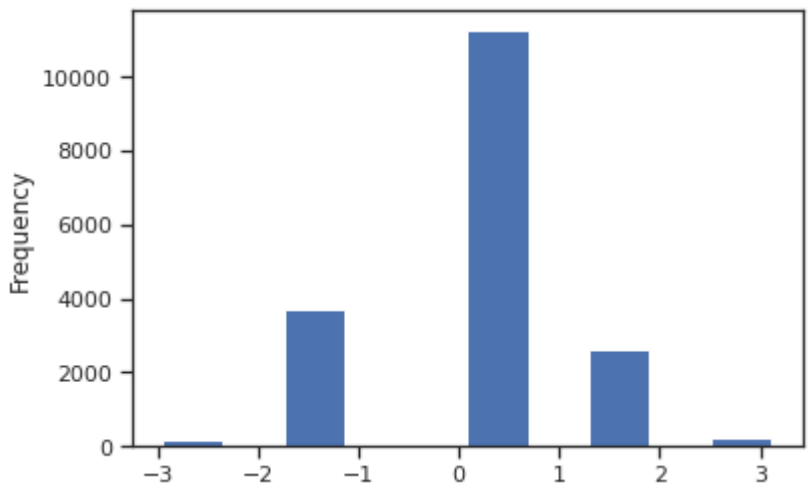
```
sc2 = StandardScaler()
df['Weak Foot Z'] = sc2.fit_transform(df[['Weak Foot']])
```

In [113]:

```
df['Weak Foot Z'].plot.hist()
```

Out[113]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f771935a128>
```



3.3 Нормализация данных

In [0]:

```
sc3 = Normalizer()
df['Weak Foot Norm'] = sc3.fit_transform(df[['Weak Foot Z']])
```

In [115]:

```
df['Weak Foot Norm'].plot.hist()
```

Out[115]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f771866e5f8>
```