دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

داده کاوی

تمرین اول

نام دانشجو: دانیال کردمدانلو

شماره دانشجویی: 9631813

1)

Unsupervised learning:

نوعی از یادگیری ماشین می باشد که در آن پترن های دیتا های تگ نشده شناسایی می شود و بر اساس آن یادگیری شکل می گیرد.

Supervised learning:

دیتا ها متناسب با pair هایشان (tag های مربوط به آن ها) به سیستم داده می شود و یادگیری بر اساس تگ های مربوطه شکل می گیرد و در واقع در reward هایی برای حدس های درست(tag از پیش تعیین شده) در نظر گرفته می شود تا سیستم یادگیری خود را بر اساس این جایزه ها شکل دهد.

Semi-supervised:

در واقع حالتی بین 2 مورد بالا می باشد و یادگیری بر اساس مقداری دیتای tag شده و تعداد زیادی دیتای خام صورت می گیرد.

Outlier:

به دیتا هایی می گ.یند که فاصله ی زیادی از نرمال دارند و نباید در بررسی ها مورد نظر قرار گیرند و این دیتا ها باید با مکانیزم های anomaly detection پیدا شوند.

Data validating:

عملیات پاکسازی دیتا و مرتب کردن آن و اطمینان از صحت دیتا را می گویند.

Data training:

مجموعه ای از دیتا اولیه که قرار است برای یادگیری, بررسی مورد استفاده قرار گیرد.

Testing data:

مجموعه ای از دیتا که جدا از دیتای train می باشد و بعد از آماده سازی سیستم جهت بررسی عملکرد به سیستم داده می شود تا خروجی مورد نظر بررسی شود.

Data warehousing:

یک سیستم برای نگه داری دیتا های جمع آوری شده و برای آنالیز و بررسی دیتا و برای ساخت گزارش دیتا های آنالیز شده مورد استفاده قرار می گرد.

Missing values:

به دیتا هایی که به صورت رندوم, اشتباهی و یا اصلا وارد نشده اند می گویند.

Independent values:

در صورتی که بین 2 متغیر هیچ رابطه ای وجود نداشته باشد و نتوان آن 2 را به شکل تابعی از دیگری نوشت.

Dimensions:

هر داده ی موجود دارای تعدادی ویژگی می باشد که به تعداد ویژگی هایی که یک داده دارد بعد یا dimension آن داده گفته می شود.

(2

برای این منظور می توان از روش های Linear Discriminant Analysis (LDA), Autoencoder و Missing Values Ratio و Low Variance Filter و... استفاده کرد.

Low Variance Filter: در این روش در صورتی که بعد از normalization ستونی از دیتا پیدا شود که واریانسی کمتر از مقدار مشخصی داشته باشد (یعنی پراکندگی دیتا ی کمی دارد) در نتیجه اطلاعات کمی را منتقل می کند. برای مثال ممکن است دیتایی وجود داشته باشد که از بین 4 مولفه ای که می تواند بگیرد در اکثر اوقات یک مولفه ی به خصوص دارد, در نتیجه جدا از

دیتا های دیگر این ستون دیتایی تقریبا همیشه یکسانی دارد پس نیازی نیست در بررسی ها به این ستون توجهی داشت.

یکی از اهداف دیتا کاوی پیدا کردن ویژگی های پنهان داده ها(extraction) بر اساس روابط یک سری ویژگی از قبل پیدا شده و یا ابتدایی می باشد(selection).

3)

Recall: در واقع مشخص می کند از کل دیتایی که باید پیدا می شد چند درصد آن یافت شده

یا معادلا : recall = TP/(TP+FP)

Precision: مشخص می کند چند درصد دیتای یافت شده اطلاعات صحیح است. (مد نظر ما)

معادلا: Precision = TP/(TP+FN)

F-score: فرمولی می باشد که بر اساس 2 معیار بالا شکل گرفته و درواقع ترکیبی از این 2 را به ما نشان می دهد که معادل 2 برابر ضرب معیار اول در دوم تقسیم بر مجموعشان.

4)

به این معنا می باشد که رابطه ی خطی بین 2 متغیر وجود ندارد. خیر متفاوت است زیرا ممکن است رابطه ی غیر خطی داشته باشند.

5)

Data cleaning: به فرایند پر کردن با مقدار دیفات به جای null value ها, حذف duplicate value ها, تعمیر و یا حذف برخی دیتا ها بر اساس ورودی های رندوم و اشتباه و... را می گویند.

Data integration: ترکیب محموعه ای از داده ها برای جلوگیری از ناسازگاری.

Data transformation: به فرایند تبدیل ساختار یا فرمت یک داده به ساختار یا فرمت دیگری data transformation می گویند.

6)

نان:N - الویه:O - پنیر:P – کره: K – مربا:M

C1 = {N=66%, O=33%, P=33%, K=50%, M=66%}

→ L1 = {N, O, P, K, M}

C2 = {(N,O) = 33%, (N,P) = 16%, (N,K) = 33%, (N,M) = 33%,

(O,P) = 16%, (O,K) = 0%, (O,M) = 0%, (M,P) = 16%, (M,K) = 50%}

→ L2 = {(N,O), (N,K), (N,M), (K,M)}

C3 = {(N,O,K) = 0%, (N,O,M) = 0%, (N,M,K) = 33%}

→ L3 = {(N,M,K) }

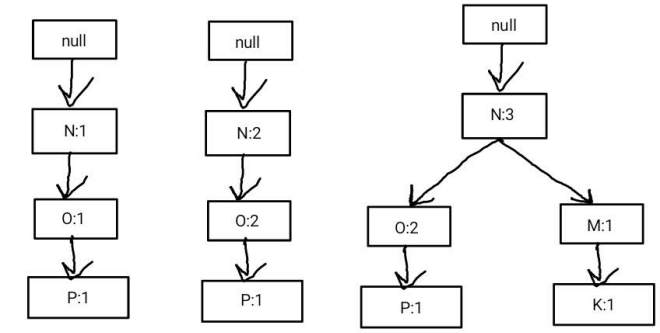
قوانین نهایی با توجه به آستانه اطمینان:

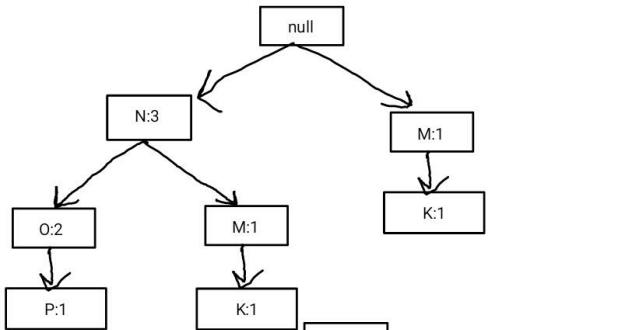
(N,M) → K :100%, (K,N) → M : 100%,

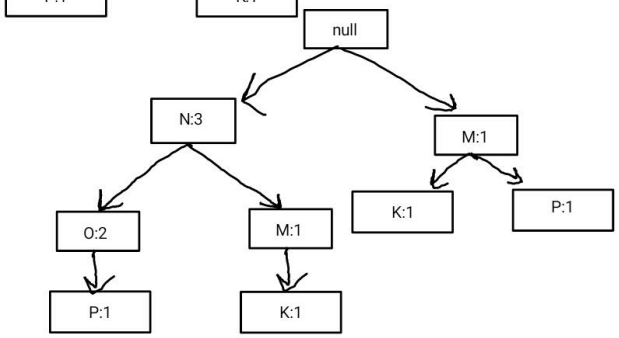
K → (N,M) : 66%, (K,M) → N : 66%

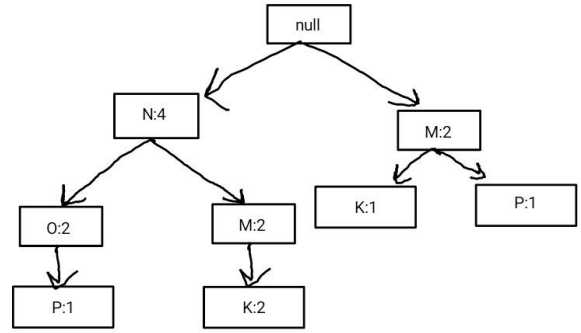
7)

الف)









ب)

33% * 6 = 2 → آیتم های پر تکرار →

{(N,O):2, (N,M):2, (N,M,K):2, (M,K):2, (N,K):2, (N,M):2}

قسمت پیاده سازی:

پیش پردازش:

1)

```
C:\Danial\University\Semester 8\Data Mining\HWs\HW1>python main.py
head and tail of data
       ID              Name                     FullName  Age  ...  LBRating  CBRating RBRating GKRating
0  158023          L. Messi                  Lionel Messi   33  ...        65        55       65       22
1   20801  Cristiano Ronaldo  C. Ronaldo dos Santos Aveiro   35  ...        64        57       64       23
2  200389          J. Oblak                   Jan Oblak   27  ...        35        36       35       92
3  192985       K. De Bruyne               Kevin De Bruyne   29  ...        78        72       78       24
4  190871          Neymar Jr    Neymar da Silva Santos Jr.   28  ...        65        52       65       23

[5 rows x 90 columns]
           ID           Name           FullName  Age  Height  ...  RWBRating LBRating CBRating  RBRating  GKRating
19015  257371    M. Nzongong      Mike Nzongong   19     179  ...         44       42       40        42        18
19016  259160        L. Bell         Lewis Bell   17     181  ...         42       41       35        41        13
19017  259157        Y. Arai         Yasin Arai   16     176  ...         45       44       39        44        17
19018  253763     R. Dinanga     Ricardo Dinanga   18     174  ...         36       34       30        34        16
19019  241493  S. Cartwright  Samuel Cartwright   19     185  ...         46       47       51        47        15

[5 rows x 90 columns]
```

2)

اسم ستون هایی که در هر ردیف دیتا ندارد(Nan) پرینت می شود.

بخشی از خروجی:

```
[18118 rows x 90 columns]
column:  4 --- ['NationalPosition' 'NationalNumber']
column:  6 --- ['NationalPosition' 'NationalNumber']
column:  11 --- ['NationalPosition' 'NationalNumber']
column:  14 --- ['NationalPosition' 'NationalNumber']
column:  16 --- ['NationalPosition' 'NationalNumber']
column:  19 --- ['NationalPosition' 'NationalNumber']
column:  20 --- ['NationalPosition' 'NationalNumber']
column:  23 --- ['NationalPosition' 'NationalNumber']
column:  24 --- ['NationalPosition' 'NationalNumber']
column:  26 --- ['NationalPosition' 'NationalNumber']
column:  33 --- ['NationalPosition' 'NationalNumber']
column:  34 --- ['NationalPosition' 'NationalNumber']
column:  35 --- ['NationalPosition' 'NationalNumber']
column:  38 --- ['NationalPosition' 'NationalNumber']
column:  40 --- ['NationalPosition' 'NationalNumber']
column:  41 --- ['NationalPosition' 'NationalNumber']
column:  43 --- ['NationalPosition' 'NationalNumber']
column:  48 --- ['NationalPosition' 'NationalNumber']
column:  50 --- ['NationalPosition' 'NationalNumber']
column:  51 --- ['NationalPosition' 'NationalNumber']
column:  53 --- ['NationalPosition' 'NationalNumber']
```

3)

```
------weights------
mean : 75.05241850683491
max : 110
min : 50
```

4)

بر اساس nation دسته بندی می شود و در نهایت با اضافه کردن یک ستون به اسم count و سورت کردن بر اساس آن دیتا را نمایش می دهیم.

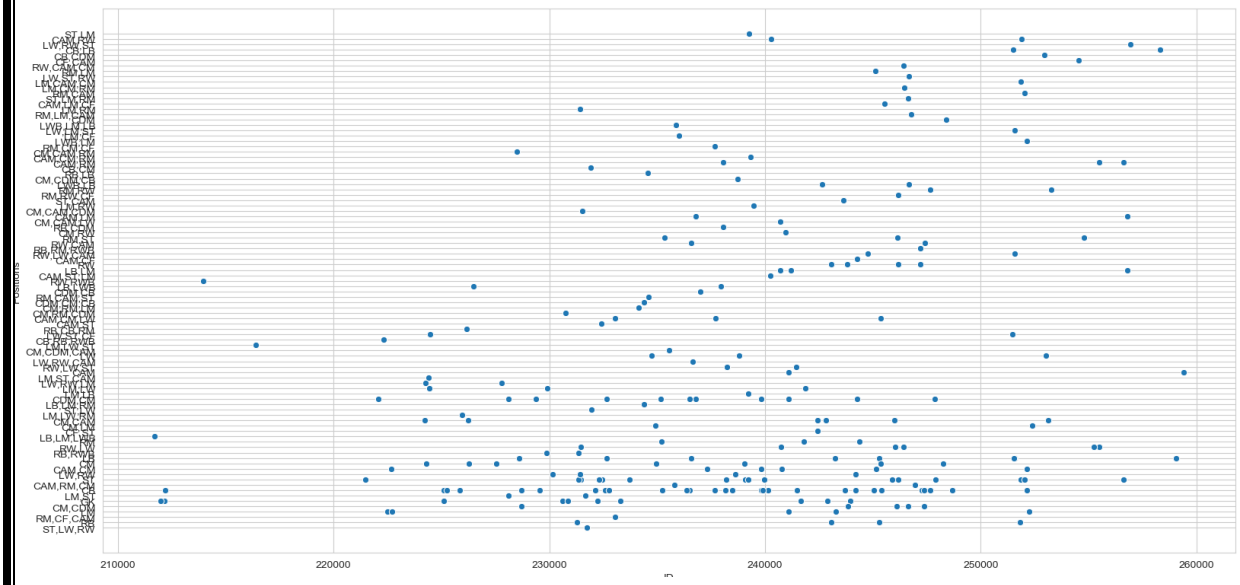
```
------players-------
Max values:
     Nationality  count
47       England   1706
58       Germany   1190
140        Spain   1084
54        France   1008
6      Argentina    945
------------------------------
Min values:
         Nationality  count
135       Singapore      1
108  New Caledonia      1
129     Saint Lucia      1
127          Rwanda      1
73        Indonesia      1
------------------------------
```

(5

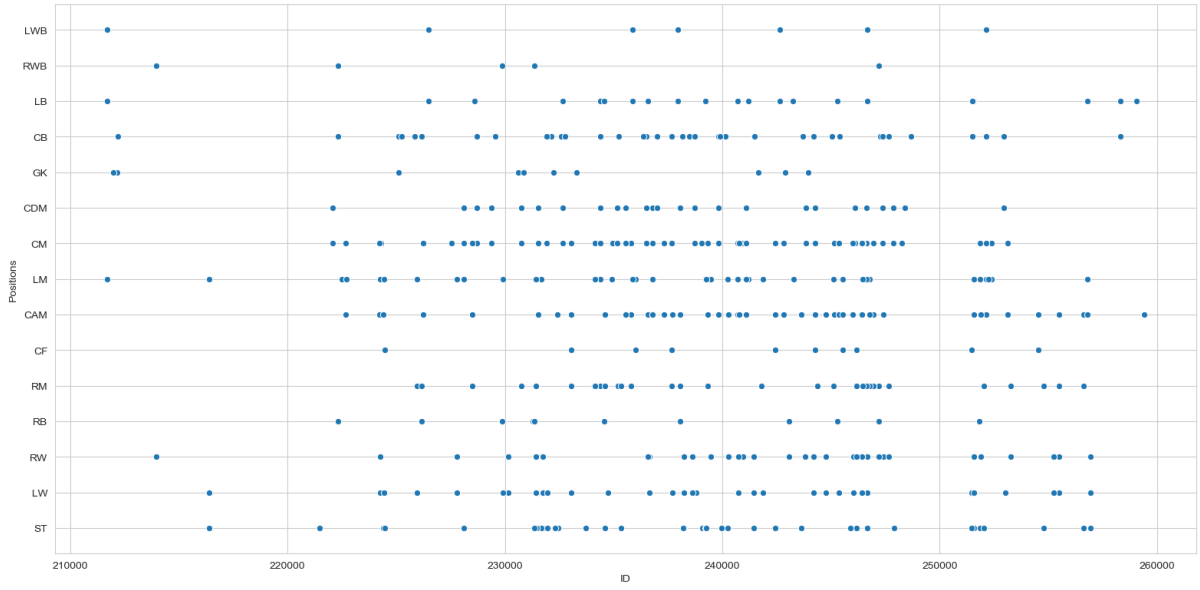
```
----Protential & Growth-----
       ID              Name             FullName  Age  Height  Weight                                PhotoUrl  ... LWBRating  CDMRating  RWBRating  LBRating  CBRating  RBRating  GKRating
12  231747          K. Mbappé         Kylian Mbappé   21     178      73  https://cdn.sofifa.com/players/231/747/21_60.png  ...        70         66         70        66        58        66        21
45  231281  T. Alexander-Arnold  Trent Alexander-Arnold   21     180      69  https://cdn.sofifa.com/players/231/281/21_60.png  ...        86         85         86        85        79        85        21
46  233049          J. Sancho          Jadon Sancho   20     180      76  https://cdn.sofifa.com/players/233/049/21_60.png  ...        69         65         69        64        52        64        22
68  222492           L. Sané           Leroy Sané   24     184      75  https://cdn.sofifa.com/players/222/492/21_60.png  ...        66         61         66        62        54        62        20
70  228702         F. de Jong        Frenkie de Jong   23     180      74  https://cdn.sofifa.com/players/228/702/21_60.png  ...        85         86         85        84        80        84        21

[5 rows x 90 columns]
          ID            Name             FullName  Age  Height  Weight                                PhotoUrl  ... LWBRating  CDMRating  RWBRating  LBRating  CBRating  RBRating  GKRating
10650  251873        Y. Demir          Yusuf Demir   17     177      65  https://cdn.sofifa.com/players/251/873/21_60.png  ...        48         43         48        43        35        43        17
12472  247649  J. Branthwaite    Jarrad Branthwaite   18     193      70  https://cdn.sofifa.com/players/247/649/21_60.png  ...        58         61         58        61        65        61        18
12798  256781        L. Netz           Luca Netz   17     188      74  https://cdn.sofifa.com/players/256/781/21_60.png  ...        62         56         62        63        62        63        18
14022  259419        T. Nakai       Takuhiro Nakai   16     178      62  https://cdn.sofifa.com/players/259/419/21_60.png  ...        49         50         49        47        42        47        16
14290  258315    B. Arrey-Mbi  Bright Akwo Arrey-Mbi   17     187      76  https://cdn.sofifa.com/players/258/315/21_60.png  ...        58         59         58        60        63        60        17

[5 rows x 90 columns]
```

6)

از آنجا که دیتا های ستون positions به صورت اری می باشند باید از ساختار رشته ای که به صورت دیفالت دارند به حالت تک المان تغییر دهیم که باعث می شود row 1 چند بار تکرار شود ولی هر بار با یکی از دیتا هایی که positions به آن اشاره دارد (برای اینکار از explode استفاده می شود) در صورت نادیده گرفتن این مورد با مشکل زیر مواجه نشویم(بسیاری از دیتا های positions تکراری است ولی چون به صورت اری کنار هم قرار گرفته دیتای جدا تلقی می شود)



بعد از اعمال اصلاحات:(نمودار نهایی پلیر ها(ID) متناسب با موقعیتشان در بازی نشان می دهد که از دیتای سوال 5 برای این مورد استفاده شده است)



7)

بر اساس دیتای بدست آمده از سوال 5 بر اساس club دسته بندی می کنیم و تعداد را بدست می آوریم و در نهایت بر اساس تعداد بدست آمده سورت می کنیم.

```
Best stars in future with Clubs in order:
             Club  count
76   Sporting CP     10
61   Real Madrid      9
13       Chelsea      8
21  FC Barcelona      8
2        Arsenal      7
```

8)

دیتا هایی از آینده داران که کلاب پلسی دارند را پیدا می کنیم و ستون ارزششان را با هم جمع می کنیم.

```
-----Chelsea-----
Chelsea Stars in future :  293900000
```

9)

داده هایی را که مقدار ستون قراردادشان 2021 می باشد و مقدار ستون NationalTeam شان نیز برابر Not in team می باشد را پیدا می کنیم و در نهایت تعدادشان را خروجی می دهیم.

```
-----retirements-----
players retirement numbers untill 2021 =  6727
```

10)

داده های مربوط به مهدی طارمی با شرط اینکه قراردادش بعد از سال 2020 تمام شود را پیدا می کنیم.

```
-----taromi data-----
          FullName Positions  WageEUR      Club
1017  Mehdi Taremi     ST,CF    16000  FC Porto
1113  Mehdi Taremi     ST,CF    16000  FC Porto
```