پروژه داده‌کاوی

مینا افضلی – زهرا رستمی

فاز دوم

بخش اول– استخراج الگوهای مکرر

در این بخش بعد از تبدیل داده ها به دسته ها یا category **(درصورت نیاز)عملیات الگویابی را انجام دادیم.**

1- **الگو مکرر ارتباط بین console و genres و publisher وdeveloper**

ابتدا بعد از بررسی دیتاست جداول transaction **را ساخته (در تصویر زیر نمونه‌ای از سطرهای اولیه‌اش آمده است) و آنها را در یک فایل** csv. **ذخیره می‌کند.**

سپس با بررسی مجموعه تراکنش ها تمامی item_set **هارا با توجه به مقدار مشخص شده‌ی** min_support **که برابر با** 0.1 **در نظر گرفته شده ساخته و در این عملیات آن مقادیری که ساپورت کمتر از حداقل دارند را در نظر نمیگیریم و در نهایت نتایج** item_set **ها را هم در یک فایل .**csv **دیگری ذخیره می کند.**

FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE

itemsets	support
Action	0.13367
Misc	0.14534
PC	0.19709
Unknown	0.19372

سپس با بررسی این مقادیر با در نظر گرفتن min_support **مورد بررسی هر کدام از** item_set **ها با بالاترین** support **را میتوان به عنوان الگوی پرتکرار پیدا شده در این بررسی برگردانیم.**

در این لیست که بررسی شده، بیشترین تکرار PC **با** support = 19.7% **است اما با توجه به این که این مقدار از** min_support **مورد قبول ما که** 50٪ **هست (با توجه به این که اگر حداقل ساپورت را نداشته باشیم توصیه شده است** 50 **درصد آیتم ست هارا در نظر بگیریم) کمتر است الگوی پرتکراری پیدا نشده است.**

2- **الگو مکرر ارتباط بین genres و console**

با عملیاتی مشابه قبل جداول زیر را ساخته به بررسی آنها می پردازیم.

Amig	All	Adventure	Action-Adv	Action	Aco
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

itemsets	support
Action	0.13367
Misc	0.145339
PC	0.197091

در این لیست که بررسی شده، بیشترین تکرار PC با support = 19% است با توجه به این که این مقدار از min_support مورد قبول ما که 50٪ هست کمتر است، الگوی پرتکراری پیدا نشده است.

3- **الگو مکرر ارتباط بین genres و publisher**

1 bit stud	10tons Ltd	10tons	10TACLE	100 Gates
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE

itemsets	support
Action	0.13367
Misc	0.145339
Unknown	0.138122

در این لیست که بررسی شده، بیشترین تکرار Misc با support = 14% است با توجه به این که این مقدار از min_support مورد قبول ما که 50٪ هست الگوی پرتکراری پیدا نشده است.

4- **الگوی مکرر بین genres و total_sales**

Education	Board Gan	Adventure	Action-Adv	Action
FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE

itemsets	support
Action	0.13367
Misc	0.145339
high_sales	0.704574
low_sales	0.294864
Misc, high_sales	0.114034

با توجه به در نظر داشتن min_support = 50% **در این مورد پس از بررسی در می یابیم، الگوی پرتکرار** high_sales **با** support = 70% **هست.**

5- **الگوی مکرر بین publisher و developer و console**

FALSE	FALSE	FALSE
FALSE	FALSE	FALSE
FALSE	FALSE	FALSE
FALSE	FALSE	FALSE
FALSE	FALSE	FALSE
FALSE	FALSE	FALSE
FALSE	FALSE	FALSE
FALSE	FALSE	FALSE
FALSE	FALSE	FALSE
FALSE	FALSE	FALSE

itemsets	support
PC	0.19709
Unknown	0.19372

در این لیست که بررسی شده، بیشترین تکرار PC **با** support = 19.7% **است با توجه به این که این مقدار از** min_support **مورد قبول ما که 50٪ هست الگوی پرتکراری پیدا نشده است.**

6- **الگو مکرر بین total_sales و developer و publisher و genre**

FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE

itemsets	support
Action	0.13367
Misc	0.145339
Unknown	0.193717
high_sales	0.704574
low_sales	0.294864
Misc, high_sales	0.114034
Unknown, high_sales	0.183876

با توجه به در نظر داشتن min_support = 50% **در این مورد پس از بررسی در می یابیم، الگوی پرتکرار** high_sales **با** support = 70% **هست.**

7- **الگو پرتکرار بین total_sales و genres و console**

7800	5200	3DS	3DO	2600
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE

itemsets	support
Action	0.13367
Misc	0.145339
PC	0.197091
high_sales	0.704574
low_sales	0.294864
Misc, high_sal	0.114034
PC, high_sale	0.172722

با توجه به در نظر داشتن min_support = 50% **در این مورد پس از بررسی در می یابیم، الگوی پرتکرار** high_sales **با** support = 70% **هست.**

بخش دوم – Clustering, Classification

خوشه بندی:

گروه بندی اولیه ما در این بخش بر اساس ویژگی های مشابه تر برای بررسی احتمالات در دسته ای که شباهات بیشتری باهم دارند به صورت 4 گروه زیر است:

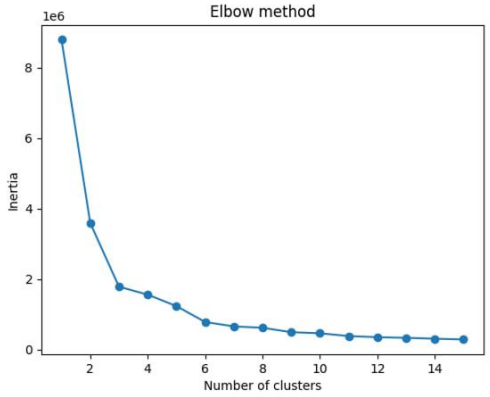
Group 0 = ['publisher_numeric', 'developer_numeric', 'console_numeric']

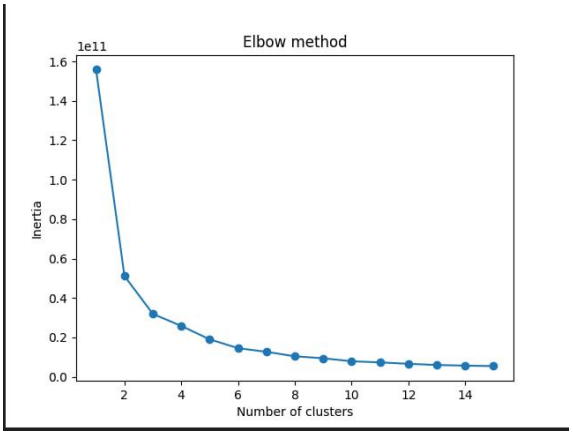
Group 1 = ['genre_numeric', 'console_numeric']

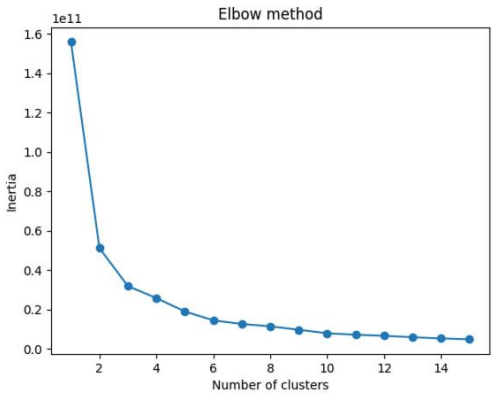
Group 2 = ['publisher_numeric', 'developer_numeric']

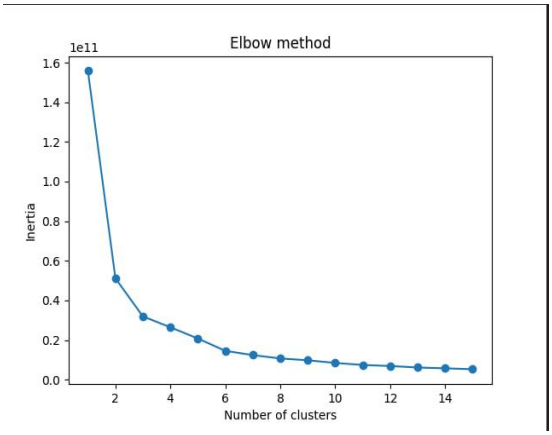
Group 3 = ['publisher_numeric', 'developer_numeric', 'console_numeric', 'genre_numeric']

با ران کردن کد نمودار هایی برای هر گروهی که در نظر گرفتیم را رسم کرده و با استفاده از Elbow Method **مقدار** k **مناسب را با بررسی این نمودار ها (محل شکست نمودار) بدست می آوریم.**



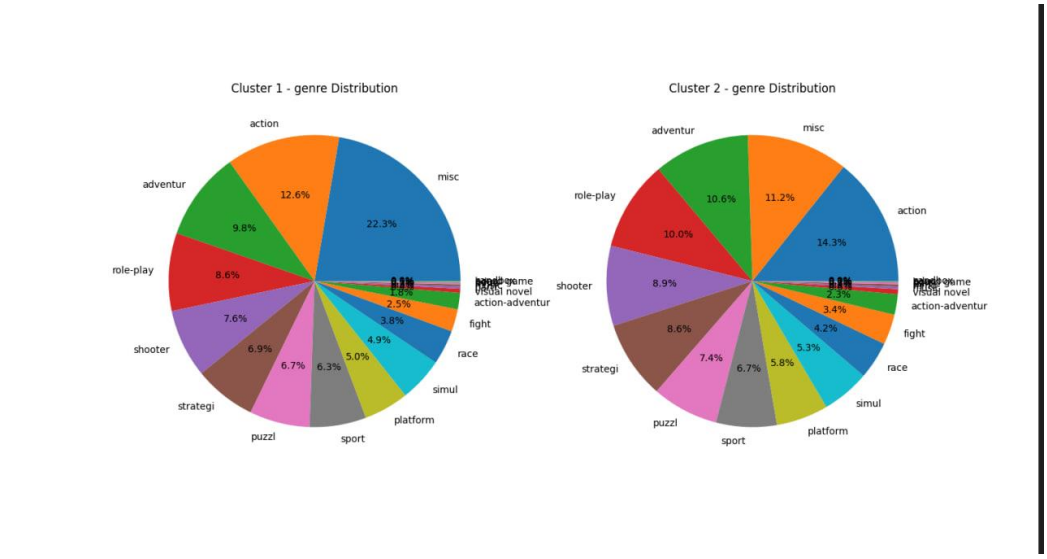


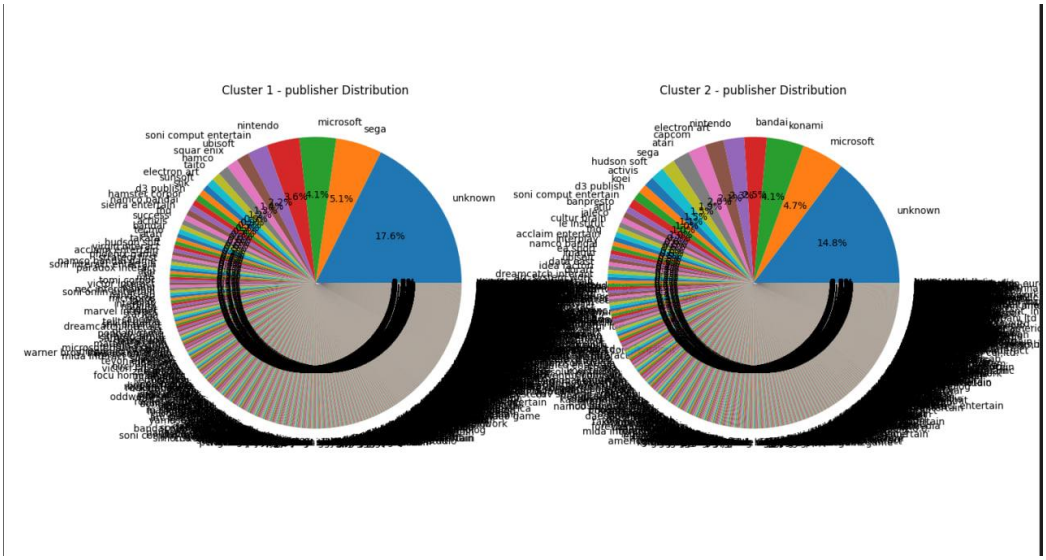


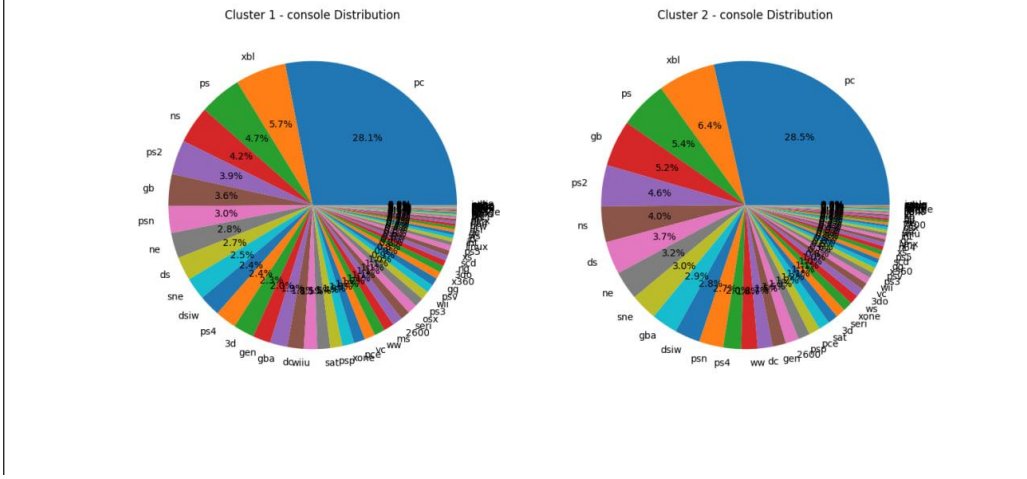


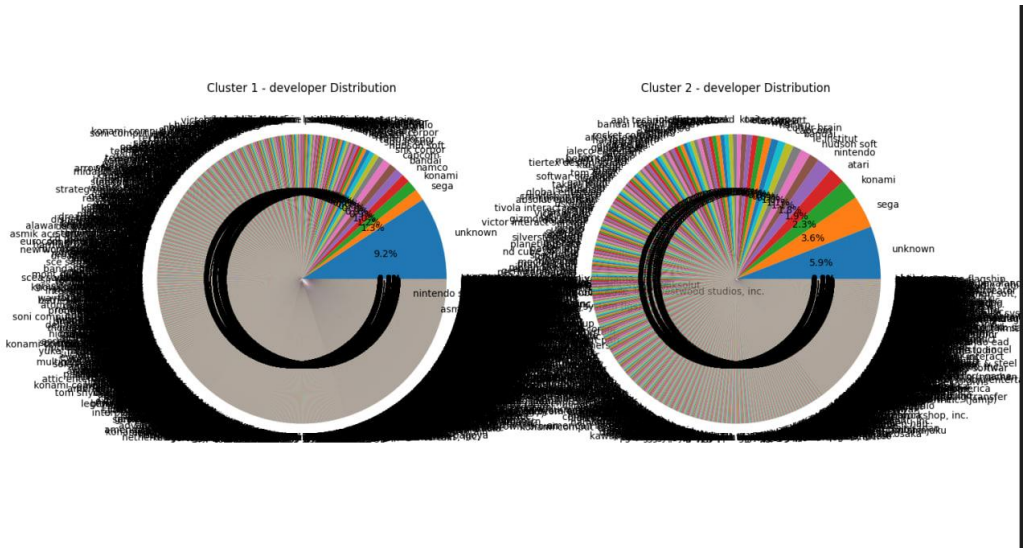
همچنین مقادیر اولیه agglomerative برای مشخص کردن k مناسب در یک فایل csv . جداگانه ذخیره می شوند.

در مرحله بعدی ما خوشه بندی ها در هر گروه را با دو خوشه بندی k-means و agglomerative بر روی نمودار هایی نمایش می دهیم و میزان فراوانی هر فیلد در آن خوشه را بررسی میکنیم که در زیر مثال هایی از این نمایش ها و مقایسه ها آمده است.






Cluster 1 - console Distribution
pc
28.1%
xbl
5.7%
ps
4.7%
ns
4.2%
ps2
3.9%
gb
3.6%
psn
3.0%
ne
2.8%
ds
2.7%
sne
2.5%
dsiw
2.4%
ps4
3d
gen
gba
dcwiiu
sat
psp
xone
pcfx
ww
ns
2600
psx
ps3
psv
wii
x360
Cluster 2 - console Distribution
pc
28.5%
xbl
6.4%
ps
5.4%
gb
5.2%
ps2
4.6%
ns
4.0%
ds
3.7%
ne
3.2%
sne
3.0%
gba
2.9%
dsiw
psn
ps4
ww
dc
gen
2600
pcfx
sat
3d
xone
ws
3do
vii


Cluster 1 - developer Distribution
nintendo
9.2%
konami
sega
bandai
namco
unknown
Cluster 2 - developer Distribution
unknown
5.9%
sega
3.6%
konami
nintendo
atari
hudson soft

طبقه‌بندی:

1- انتخاب ویژگی:

برای انتخاب ویژگی‌هایی که بیشتر در پیش‌بینی متغیری مانند user_score **نقش دارند، چندین روش وجود دارد که می‌توان از آن‌ها استفاده کرد. این روش‌ها به دو دسته کلی تقسیم می‌شوند: روش‌های مبتنی بر فیلتر** (Filter Methods) **و روش‌های مبتنی بر مدل** (Model-based Methods).

1. **روش‌های مبتنی بر فیلتر** (Filter Methods)

1. همبستگی (Correlation):

– محاسبه همبستگی بین هر ویژگی و user_score.

– ویژگی‌هایی که همبستگی بالاتری با user_score **دارند، برای مدل انتخاب می‌شوند.**

```
User Score                1.000000
Publisher_numeric         0.053307
Title_numeric             0.033315
User Ratings Count        0.027685
Developer_numeric         0.007089
Product_Rating_numeric   -0.015238
Genres_numeric           -0.017952
Genres_Splitted_numeric  -0.018356
Platforms_Info_numeric   -0.031567
Name: User Score, dtype: float64
```

2. آزمون‌های آماری:

– استفاده از آزمون‌های آماری مانند آزمون تی (T-test)**، آنالیز واریانس** (ANOVA) **یا آزمون کای دو** (Chi-square test) **برای تعیین ویژگی‌های مهم.(برای ویژگی‌های عددی از** ANOVA **و برای ویژگی‌های دسته‌بندی از** Chi-square **استفاده می‌کنیم.)**

3. اهمیت متقابل اطلاعات (Mutual Information):

– اندازه‌گیری میزان اطلاعات مشترک بین هر ویژگی و user_score.

– ویژگی‌هایی که مقدار بالاتری از اطلاعات مشترک دارند، برای مدل انتخاب می‌شوند.

2. **روش‌های مبتنی بر مدل** (Model-based Methods)

1. رگرسیون:

– استفاده از مدل‌های رگرسیون مانند رگرسیون خطی یا لجستیک برای تعیین اهمیت ویژگی‌ها.

2. **درخت تصمیم** (Decision Tree):

– **استفاده از مدل‌های درخت تصمیم و بررسی ویژگی‌های مهم بر اساس میزان کاهش عدم قطعیت** (impurity reduction).

3. **جنگل تصادفی** (Random Forest):

– **استفاده از مدل‌های جنگل تصادفی و بررسی ویژگی‌های مهم بر اساس اهمیت ویژگی** (feature importance).

4. **ماشین بردار پشتیبانی با ویژگی‌های رگرسیون** (SVR):

– **استفاده از ماشین بردار پشتیبانی** (Support Vector Regression) **و بررسی وزن‌های ویژگی‌ها.**

پس از محاسبه اهمیت هر ویژگی با یکی یا چندین روش فوق، می‌توان ویژگی‌های با اهمیت بالا را برای مدل نهایی انتخاب کرد. همچنین می‌توان از ترکیبی از این روش‌ها برای انتخاب بهترین ویژگی‌ها استفاده کرد.

2- **انتخاب مدل**

در ادامه با پیاده سازی 3 مدل Decision Tree, Random Forest, Svm **دقت ها یا** acuuracy **را برای هر مدل محاسبه میکنیم و با توجه به اینکه دقت در** svm **بالاتر هست مقادیر** Precision,recall **و** F1 measure **را محاسبه میکنیم.**

متد: SVM

روی همه ی ویژگی ها به صورت تکی svm **را اجرا می کنیم و دقت بدست آمده ی هر کدام را ذخیره می کنیم.**

در ادامه بر اساس دقت بدست اومده ویژگی هارا sort **می کنیم و به صورت تکی از آن آرایه برمی داریم تا زمانی که بیشترین** accuracy **را داشته باشیم و زمانی که دقت ما شروع به کاهش کرد دیگر انتخاب ویژگی را ادامه نداده و همان لیست را به عنوان ویژگی های انتخاب شده برگردانده و مقادیر لازم را محاسبه می کنیم.**