UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

GOOGLE PLAY APPS

Xu, Ange
Dai, Zhongkai
Campeny, Eloi
Moure, Ximena
González, Victor
Chriki, Fatima Zohra

Multivariate Analysis ◇ Dec'22

TABLE OF CONTENTS

01 Problem definition

02 Previous study

03 Clustering

04 Profiling

05 Decision Trees

06 Discriminant Analysis

07 Conclusions

01

Problem definition

Problem definition

Google Play

Analyze which factors can influence the rating of an app

02

Previous study

Main Preprocessing Tasks

FEATURE SELECTION	NEW VARIABLES DERIVATION	VARIABLES TRANSFORMATION
SEGMENTATION OF POPULATION	MISSING DATA	UNIVARIATE & MULTIVARIATE OUTLIERS

Knowledge obtained from D3

From **PCA**: 3 dimensions

- `Rating.Count`, `DaysLastUpdate` and `Installs` are the most important variables in explaining the dataset.
- Higher number of installs or votes does not mean a high rating.
- Frequency of updates does not have any effect on the rating.

From **MCA**: 5 dimensions

- Entertainment
- Procrastination
- Companionship
- Longevity
- Helpfulness in a person's lifestyle

Knowledge obtained from D3

From **MFA**:

- The older an app is, the more popular it is. We can also sometimes see that the newer apps tend to have less size and short names.
- There are no clear clusters of individuals in the data.
- In general, not all individuals are seen the same by all the groups, there is a high difference, specially between App Features and Popularity.

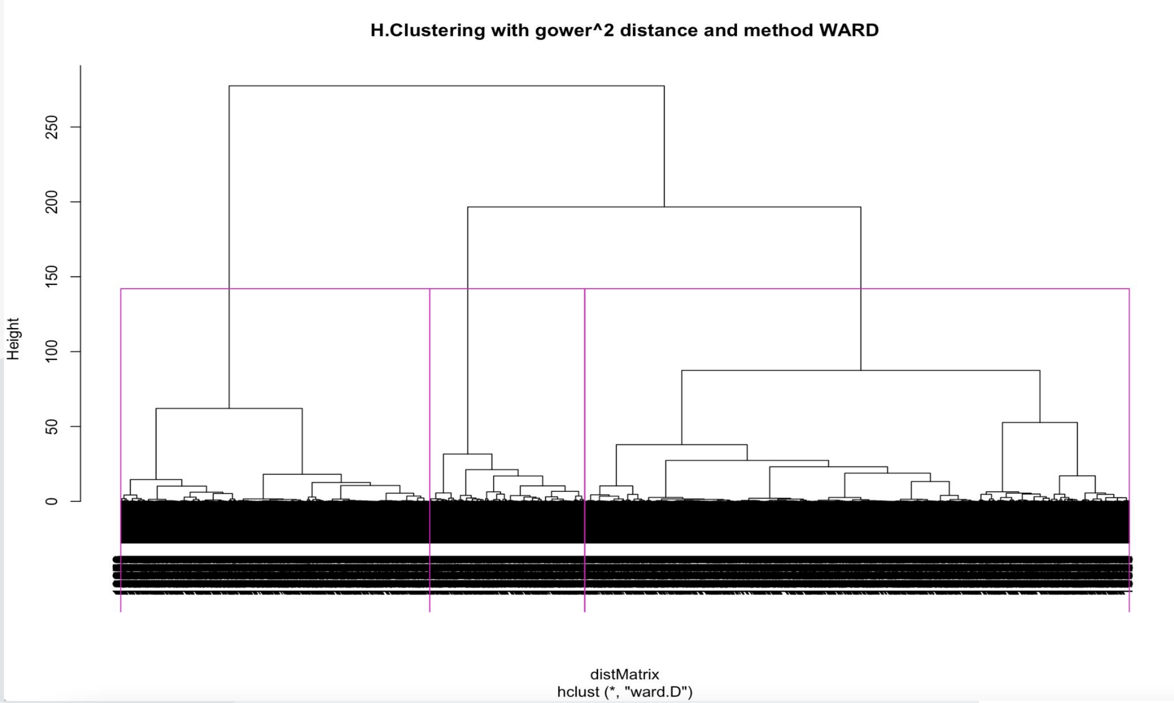
From **Association rules**:

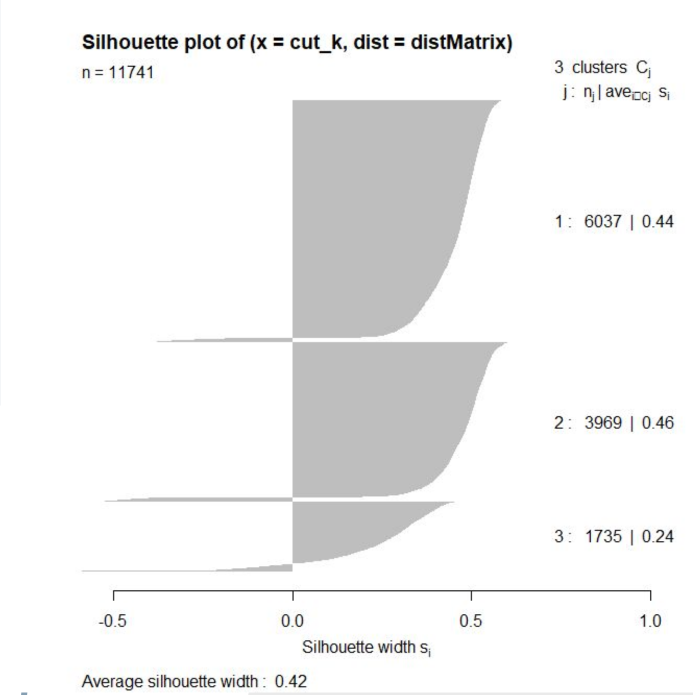
- Apps that belong to games category, have a long name or have a minimum android 4 are ad supported.
- Apps that don't have in app purchases, have ads or its release date is mid have minimum android 4.

03

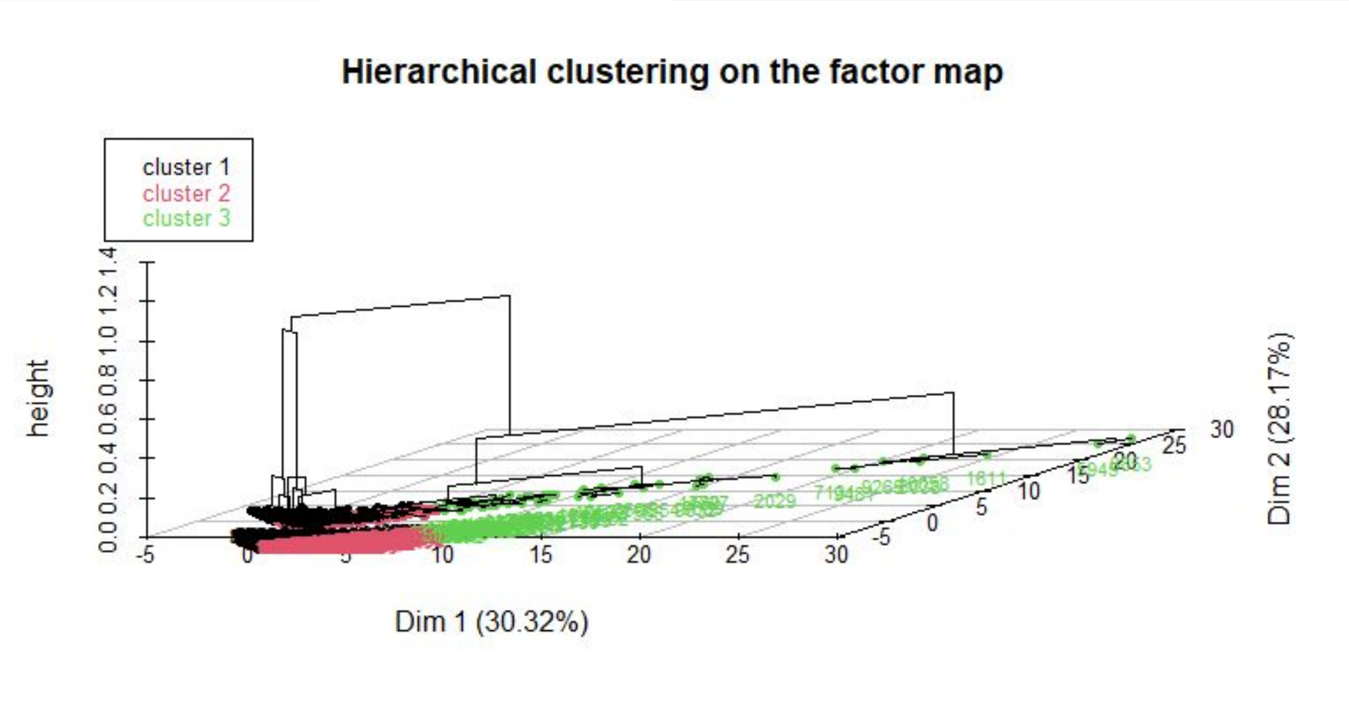
Clustering

Dendrogram and Silhouette plot





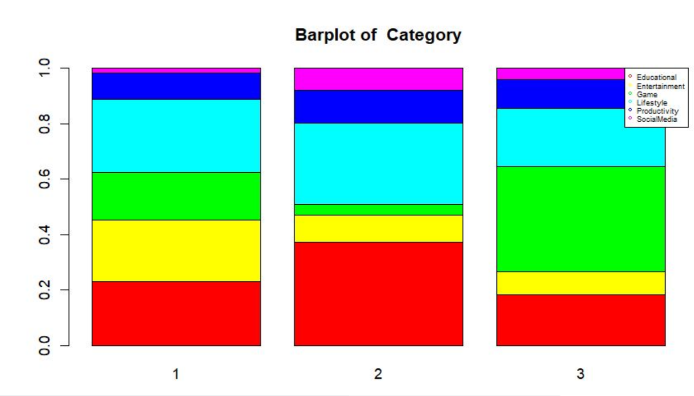
Dendrogram with HCPC

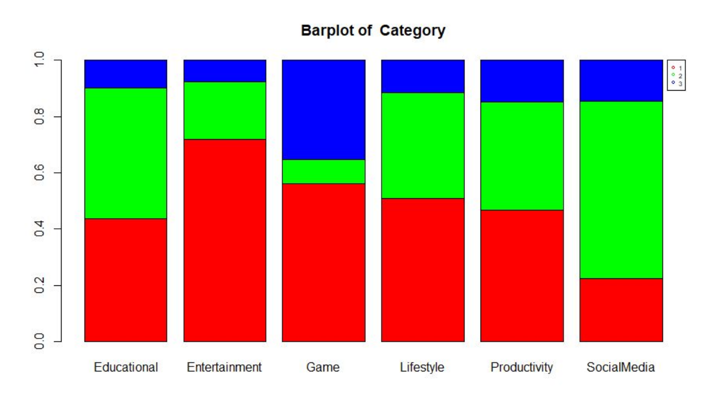


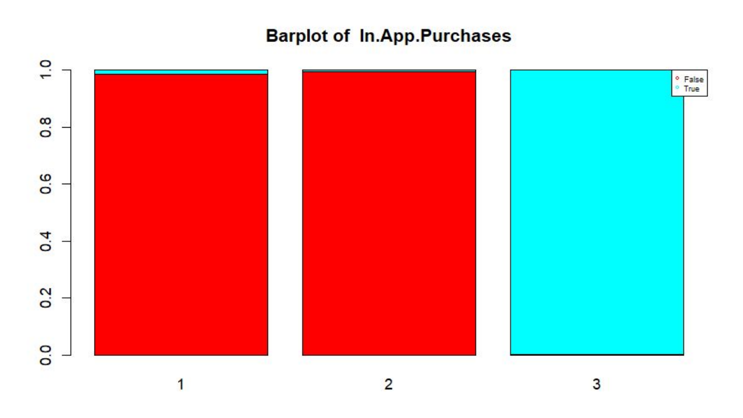
04

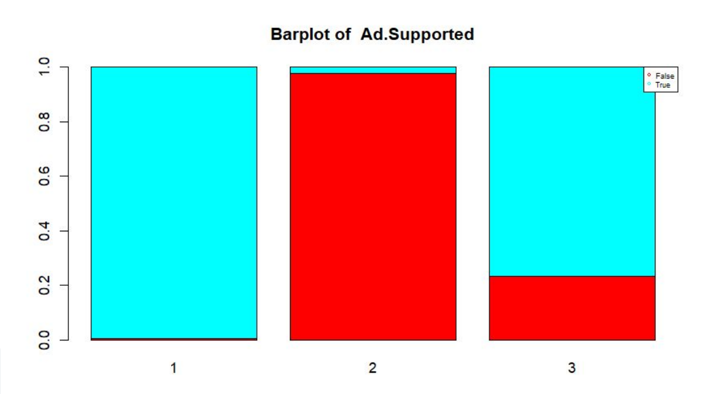
Profiling

Selection of relevant categorical variables

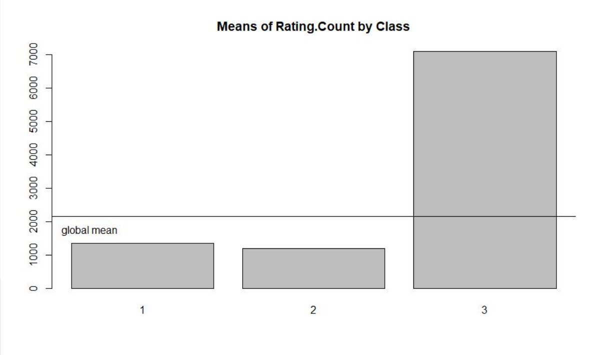


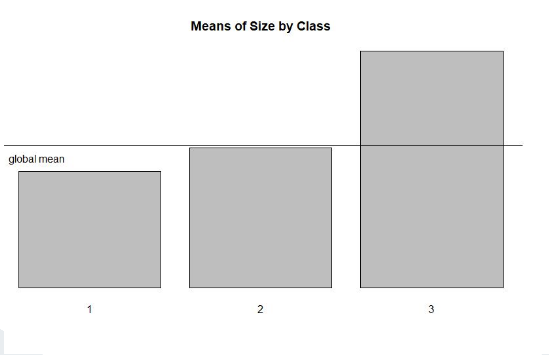


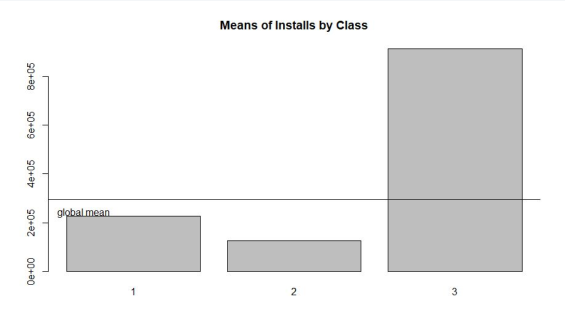


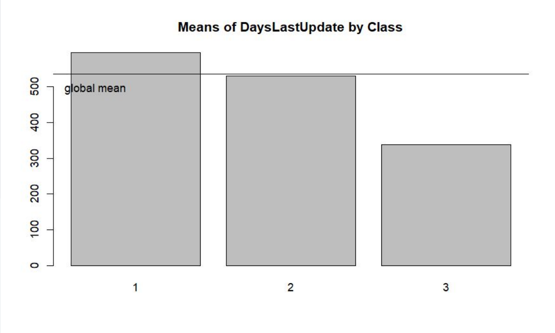


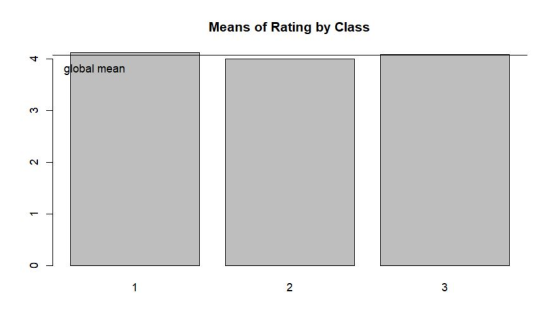
Selection of relevant numerical variables


Means of Rating.Count by Class


Means of Size by Class


Means of Installs by Class


Means of DaysLastUpdate by Class


Means of Rating by Class

Profiling of clusters

Cluster 1: Lifestyle, Entertainment & Educational
Cluster 2: Lifestyle & Educational, short names
Cluster 3: Games

Cluster	**1**	**2**	**3**
Rating.Count	⬇	⬇	⬆⬆⬆
Size	⬇	=	⬆
Installs	⬇	⬇⬇	⬆⬆⬆
DaysLastUpdate	⬆	=	⬇
Ads	✓	🚫	✓
In-app Purchases	🚫	✓	✓

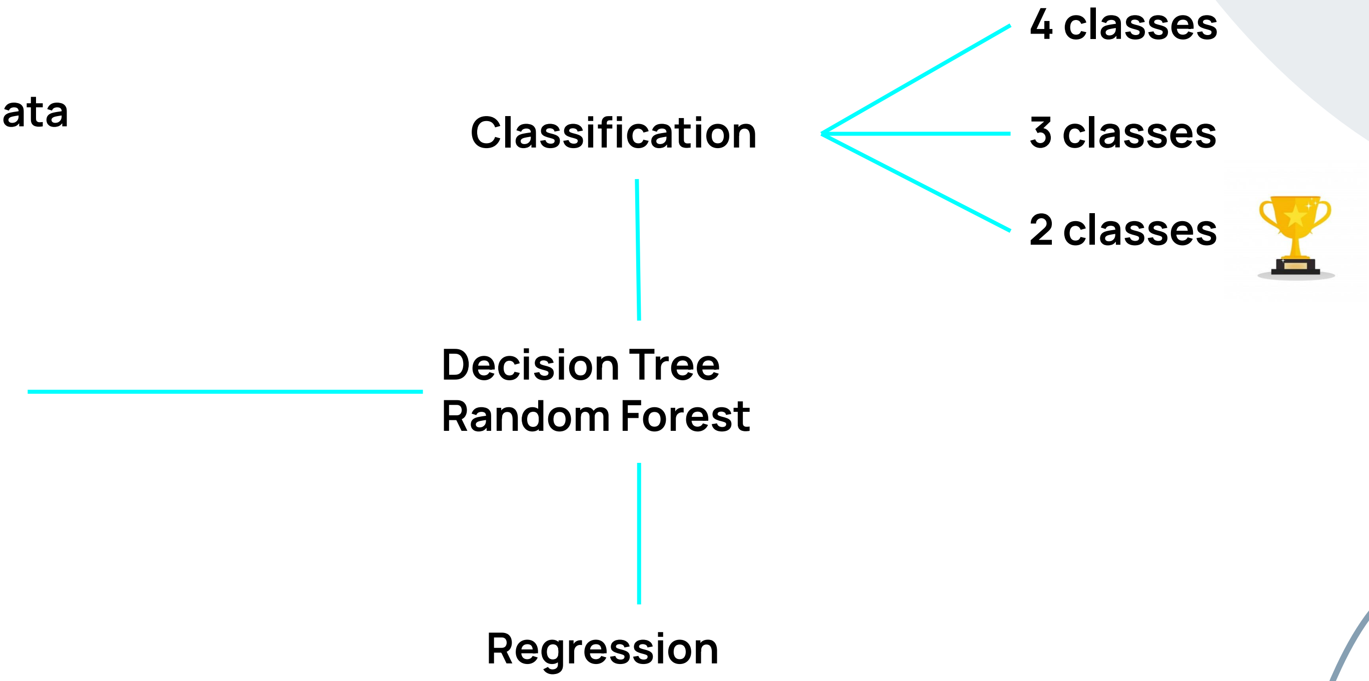
05

Decision Trees

Approaches

- Balance Data

- Generate Model



Classification

4 classes

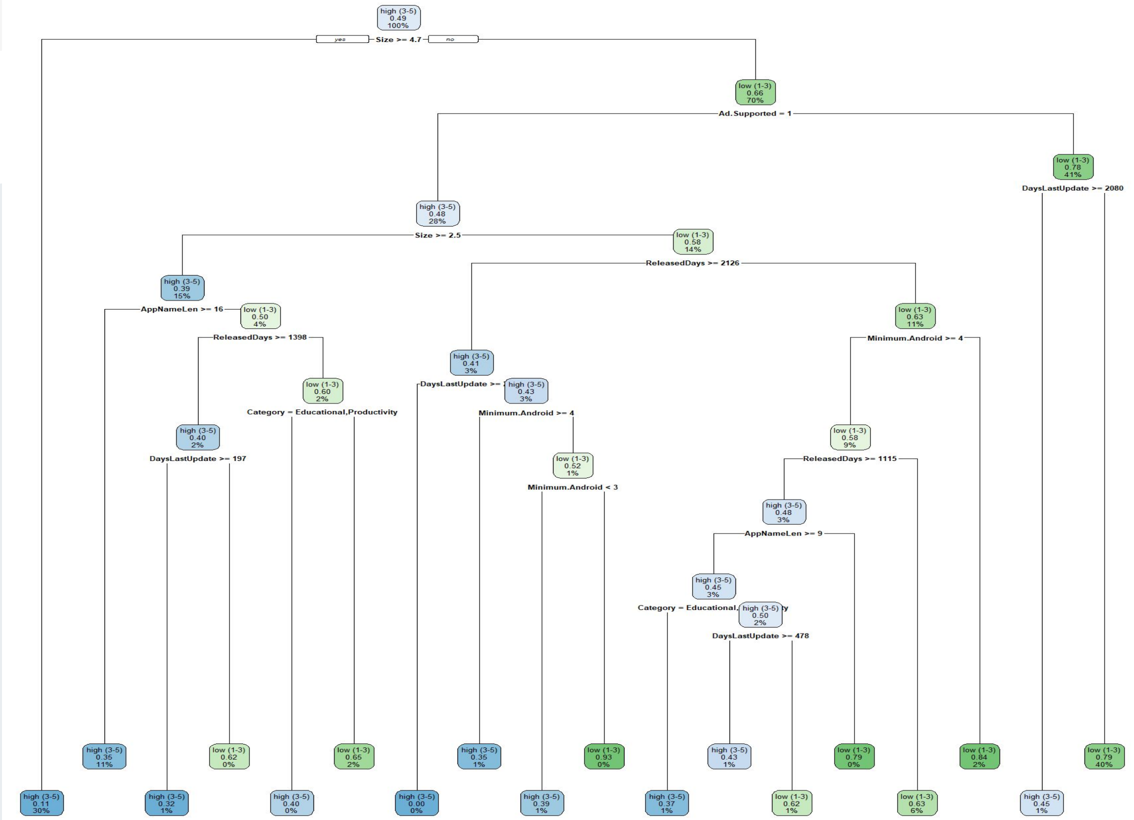
3 classes

2 classes 🏆

Decision Tree
Random Forest

Regression

Decision tree



MAIN RULES

- *Rule 1:*

  *IF **size > = 4.7***

  *THEN **rating = high***

- *Rule 2:*

  *IF **size < 4.7** &&*
  ***ad.Supported != 1** &&*
  DaysLastUpdate < 2080

  *THEN **rating = low***

  *ELSE **rating = high***

- *Rule 3:*

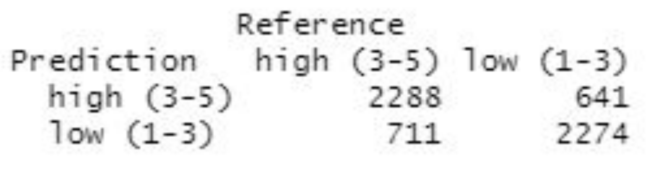
  *IF **size < 4.7** && **ad.Supported***
  ***= 1** && **Size > = 2.5** &&*
  AppNameLen > = 16

  *THEN **rating = high***

Model Evaluation

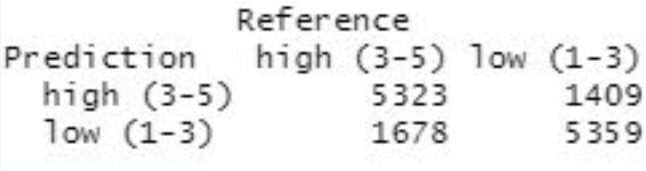
- CONFUSION MATRIX

Testing



```
                Reference
Prediction     high (3-5) low (1-3)
  high (3-5)         2288       641
  low (1-3)           711      2274
```

Training



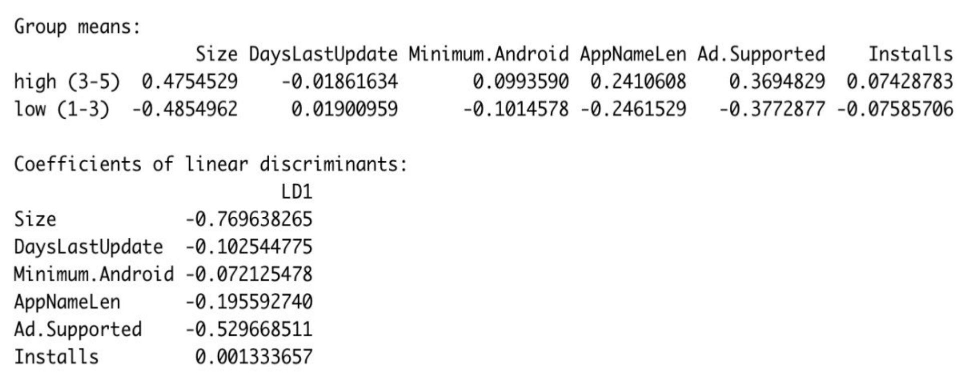
```
                Reference
Prediction     high (3-5) low (1-3)
  high (3-5)         5323      1409
  low (1-3)          1678      5359
```

- **77%** ACCURACY
- **74%** RECALL
- **79.5%** PRECISION

06

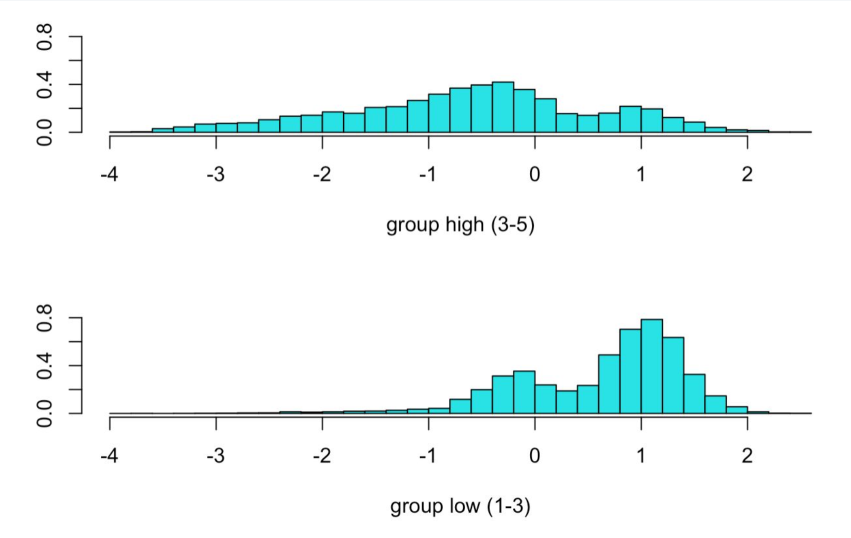
Discriminant Analysis

LDA



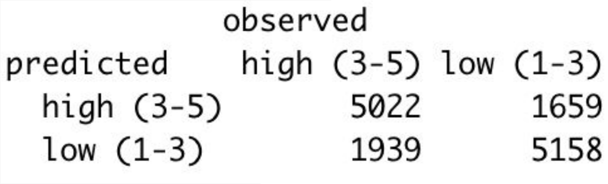
```
Group means:
                 Size DaysLastUpdate Minimum.Android AppNameLen Ad.Supported     Installs
high (3-5)  0.4754529    -0.01861634       0.0993590  0.2410608    0.3694829  0.07428783
low (1-3)  -0.4854962     0.01900959      -0.1014578 -0.2461529   -0.3772877 -0.07585706

Coefficients of linear discriminants:
                        LD1
Size            -0.769638265
DaysLastUpdate  -0.102544775
Minimum.Android -0.072125478
AppNameLen      -0.195592740
Ad.Supported    -0.529668511
Installs         0.001333657
```



Evaluation

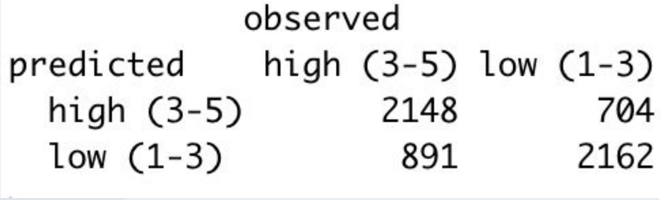
Training



```
              observed
predicted      high (3-5) low (1-3)
  high (3-5)         5022      1659
  low (1-3)          1939      5158
```

- Accuracy: 74%
- Misclassification rate: 26%

Testing



```
              observed
predicted      high (3-5) low (1-3)
  high (3-5)         2148       704
  low (1-3)           891      2162
```

- Accuracy: 73%
- Misclassification rate: 27%

07

Conclusions

Conclusions

- The rating of an app is determined by:
    - Size :: high size → high rating
    - Supports ads :: ad supported → high rating
    - Days last update :: less days → low rating
    - Name Length :: long name → high rating
- Future analysis
    - Categories games or lifestyle
    - Limit time range of apps
    - `Installs` as response variable
    - Analyze non-free apps & exceptional apps

Thank you for your attention