جامعة الملك سعود
King Saud University

Knowledge Management and Data Mining
MIS 419

DATA MINING PROJECT

Google Play data using Orange

orange

Data Mining Fruitful and fun

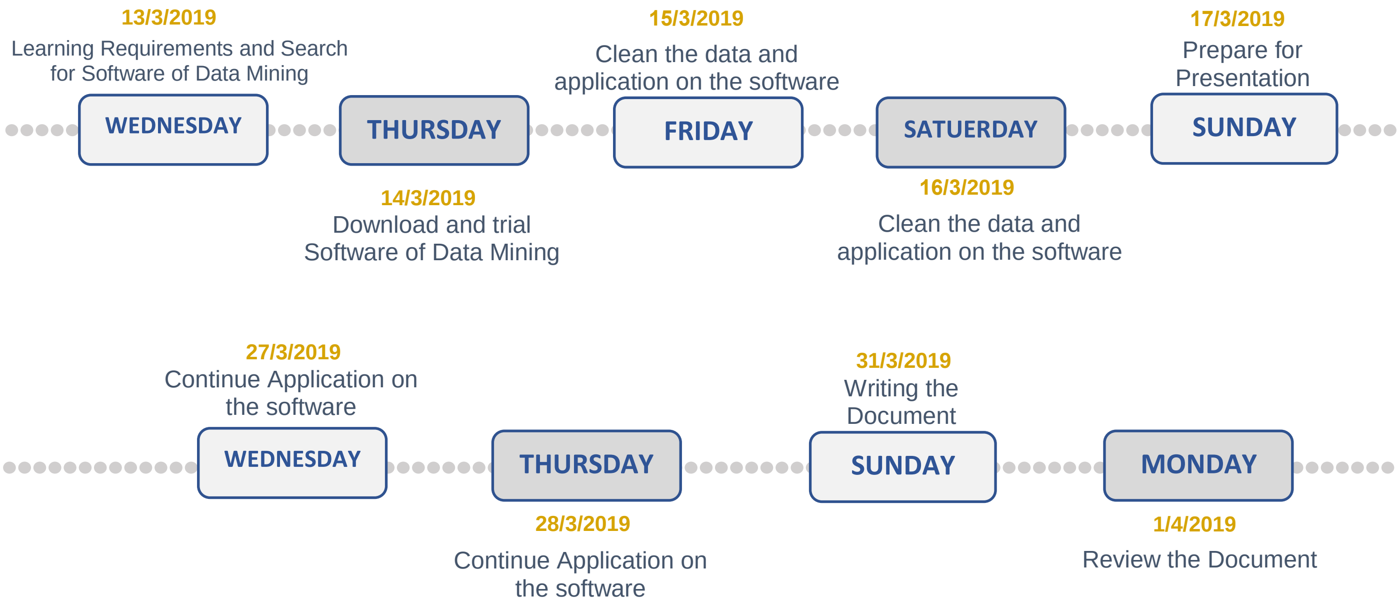
STUDENT NAME	STUDENT ID
Banan Alaqeel	436201599
Raghad Alrumaih	436202314
Rahaf Alsalamah	436200706
Sara Alkhudair	436201596

Group : 2
Section: 54771
Lecturer: Monira Essa Aloud
Semeter : 2nd Semester 2019

Table of Content

Table of Figure

Time Line



13/3/2019
Learning Requirements and Search for Software of Data Mining
WEDNESDAY

14/3/2019
Download and trial Software of Data Mining
THURSDAY

15/3/2019
Clean the data and application on the software
FRIDAY

16/3/2019
Clean the data and application on the software
SATUERDAY

17/3/2019
Prepare for Presentation
SUNDAY

27/3/2019
Continue Application on the software
WEDNESDAY

28/3/2019
Continue Application on the software
THURSDAY

31/3/2019
Writing the Document
SUNDAY

1/4/2019
Review the Document
MONDAY

Introduction

Is this project our main objective is to solve google play case by using data mining, we conducted experiments in many software available in data mining such as orange, RapidMiner, Xlminer and Weka, before choosing the right program for us, we would like to take an overview of the programs available in this area.

In this document we will discuss the steps of data mining from data integration, data selection, data cleaning, data transformation, finally data mining where we will use the required techniques such as classification and then we will use validation experiment using Orange. In the following we will talk about this software and strengths and limitations.

Finally, in this document we will we discuss our recommendations, to help google play in solving their problem.

Goals

- Gain practical experience from data mining software, specifically RapidMiner and Orange.
- Interact with Google Play data using data mining algorithms.
- Discover hidden patterns in the Google Play Store Dataset.
- To apply what we have learned and test our knowledge.

Orange

Orange is all about data visualizations that help to uncover hidden data patterns, Orange is a great data mining tool for beginners and experts, from user interface they can focus on data analysis.

In Orange, data analysis is done by stacking components into workflows. Each component, called a widget, embeds some data retrieval, pre-processing, visualization, modelling or evaluation task. Additional widgets are available through add-ons and allow for a more focused and topic-oriented research.

○ Orange Strength

- The flexibility of the tool and the power to invent new combinations of data mining methods.
- Orange comes from the combination of visual programming and interactive visualizations.
- Enables a simple analysis even without knowing a whole lot about statistics, machine learning, or exploratory data mining in general.

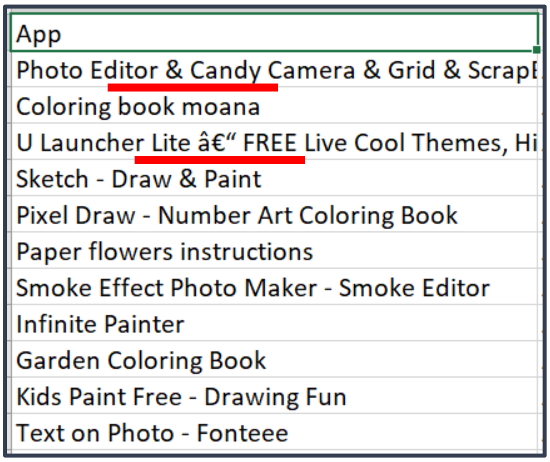
○ Orange Limitation

- Some processes require typing in Python code This is difficult for beginners or those who do not know the language.
- Some operations require adding visualize such as tree viewer.
- The feature does not support the addition of simple explanations about the tools which require visiting Orange website to know about them.
- Some visuals result is not clear without labels like some of the components in the decision tree.

Cleaning Up Google Paly Data

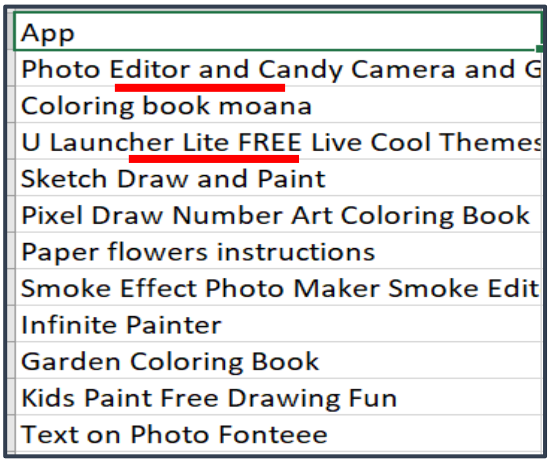
At this point, we first explored the data and noticed that there were missing values and symbols that were nonprintable or in other words incomprehensible. First we removed the rows whose values we cannot predict and complete the rows for which we can search for values such as the price of the application using Excel.

We then used the Excel filtering feature to search for unacceptable or missing values, then we uploaded the enhanced data to Orange, the next part of cleaning up the data was using the **Preprocess** widget in Orange.



App
Photo Editor & Candy Camera & Grid & Scrapl
Coloring book moana
U Launcher Lite â€“ FREE Live Cool Themes, Hi
Sketch - Draw & Paint
Pixel Draw - Number Art Coloring Book
Paper flowers instructions
Smoke Effect Photo Maker - Smoke Editor
Infinite Painter
Garden Coloring Book
Kids Paint Free - Drawing Fun
Text on Photo - Fonteee

Figure 1 Before Cleaning Data



App
Photo Editor and Candy Camera and G
Coloring book moana
U Launcher Lite FREE Live Cool Themes
Sketch Draw and Paint
Pixel Draw Number Art Coloring Book
Paper flowers instructions
Smoke Effect Photo Maker Smoke Edit
Infinite Painter
Garden Coloring Book
Kids Paint Free Drawing Fun
Text on Photo Fonteee

Figure 2 After Cleaning Data

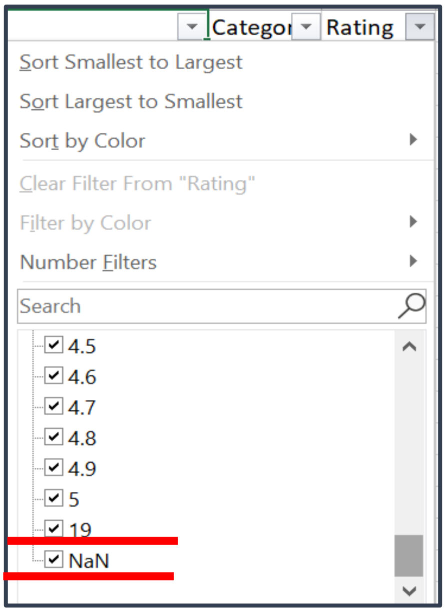


Figure 3 Example of Filtering column with missing values or an acceptable Data

The Classifier

the classifier we chose was rating, since it is affect installs the most and we have already used installs as a classifier in the decision tree that we use to control the size of the database, also Rating is the measure satisfaction of the application, and how people are using it.

Decision tree.

To control the size of the decision tree, we started by creating a decision tree presenting the apps only with more than 1000.000.000 installs, which is the largest number of installs that google play apps have. We maintain that by using the **select row** widget in orange, then we created the decision tree using the tree model.

The least rating that apps with more than 1000.000.000 installs have was 3.9. So, we start creating our decision tree based on that.

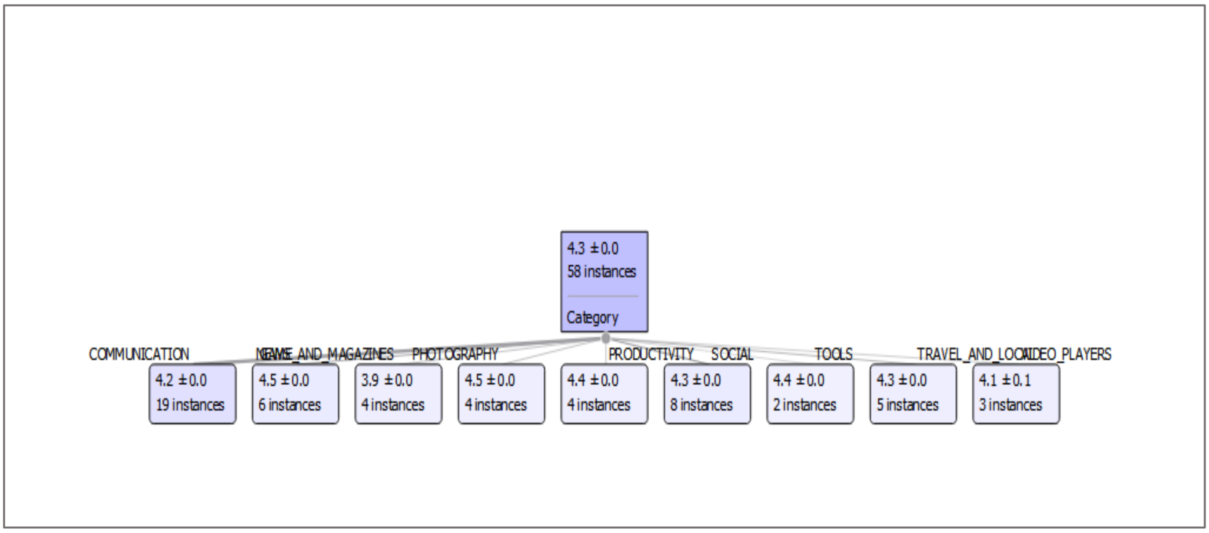


Figure 4 Decision Tree of the Category with more than 1.000.000.000 installs

The following Decision Tree represent the content rating within the chosen category that have apps with more than 1000.000.000 installs. The most installed apps fall in everyone content rating in communication, Tools, Travel and local, and fall in everyone +10 in news and magazines and Teen photography, Productivity, Video players.

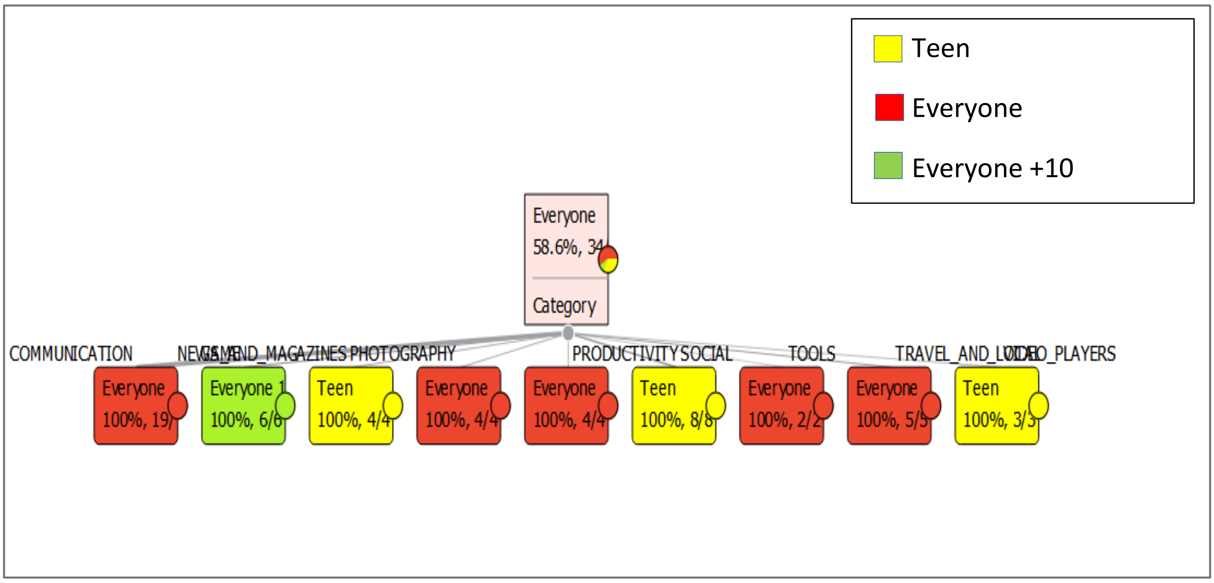


Figure 5 Decision Tree within the chosen Category

The final Decision Tree, that we create with a selected few categories, rating that is only greater than 3.9 and the selected content rating from the previous decision trees.

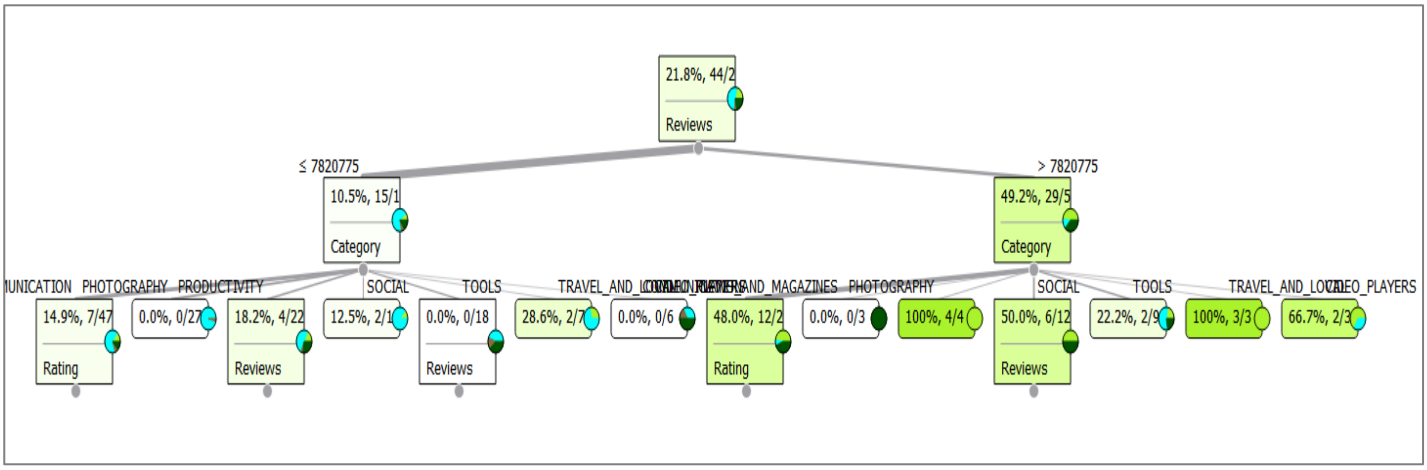


Figure 6 Final Decision Tree

The most installed apps fall within these categories (Communication, news and magazines, photography, Productivity, social, Tools, Travel and local, Video players) , have rating above 3.9, and fall with everyone content rating in communication, Tools, Travel and local, everyone +10 in news and magazines and teen photography, Productivity, Video players.

Clustering

K-means

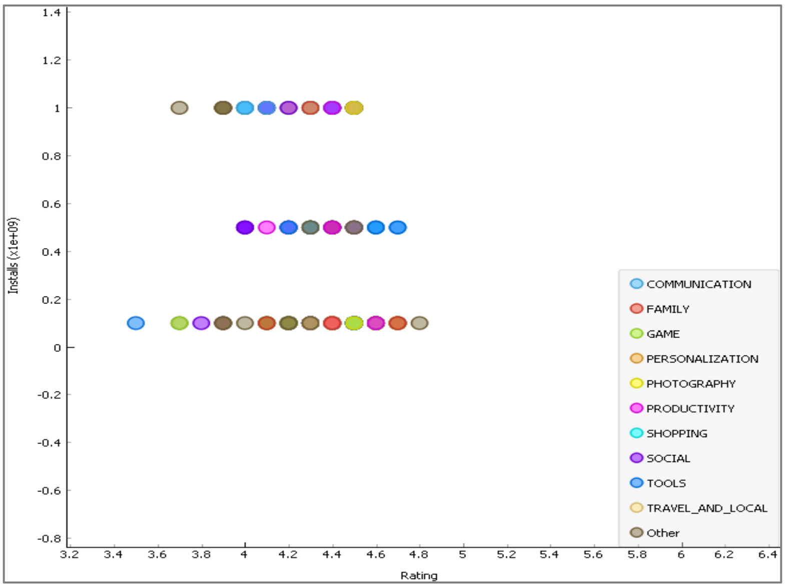
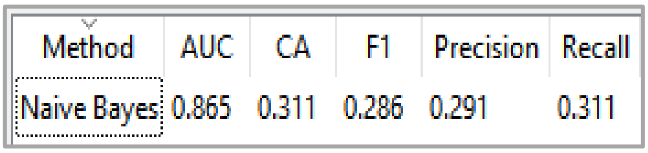


Figure 7 K-means

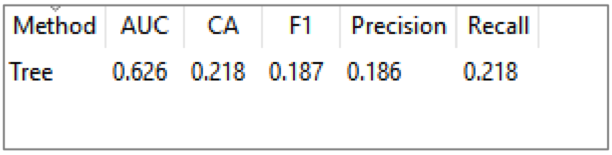
As we can see the plot show That most of the categories that has the apps with the largest number of installs have a high total rating.

Validation



Method	AUC	CA	F1	Precision	Recall
Naive Bayes	0.865	0.311	0.286	0.291	0.311

Figure 8 Naive Bayes



Method	AUC	CA	F1	Precision	Recall
Tree	0.626	0.218	0.187	0.186	0.218

Figure 9 Tree

To evaluate our models we used the **Test and score** widget in Orange it evaluate the model based on many criteria such as **AUC** (Area under ROC) which is the area under the receiver-operating curve , **CA** (Classification accuracy)which is the proportion of correctly classified examples,**F1** is a weighted harmonic mean of precision and recall ,**Precision** is the proportion of true positives among instances classified as positive and **Recall** is the proportion of true positives among all positive instances in the data. based on that

- 30% of the variables in the naïve Bayes model were classified correctly

- 20% of the variables in the decision tree model were classified correctly

Recommendations.

- Most of the apps with the largest installed number fall under communication so we advise google play to develop more apps under this category.

- The most installed apps have a rating of 3.9 and above there for rating is a good indicator if the app will be installed or not.

- 75% of the most installed apps in communication have more than 3419513 reviews therefore it influences the success of an app so google play must focus on encouraging customers to review apps and make sure the reviews are positive.

- In each category the most installed apps fall under some content rating ,so we advice google play to focus on targeting those customer segments when developing an app on those categories

Conclusion

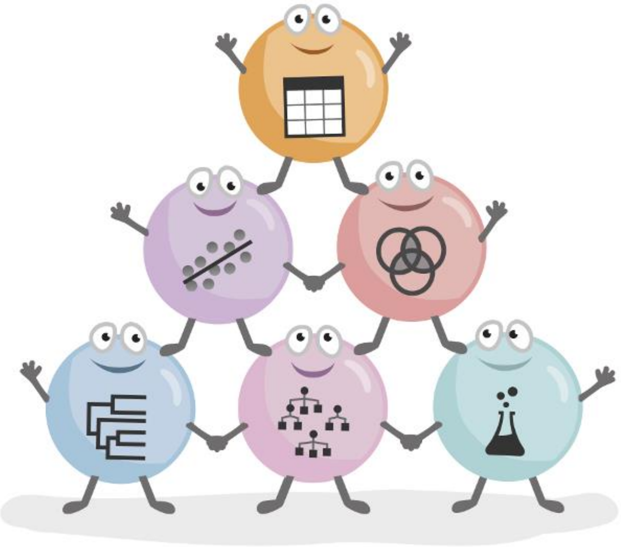
At the conclusion of this project we want to note that we have acquired practical experience that we seek to develop in the future, and also implementing our knowledge helps us to find our weakness and overcome them , we also know have a knowledge that we would've never get without doing these project it has been one of the most important and valuable experience to us .

Lesson learned

- The most important steps of data mining are the data exploration stage, from which we can learn the results we want to extract.
- Data mining software has different characteristics but all of them perform the same purpose and results if they are used correctly.

Appendix

Date & time of meeting	Place of meeting	Group members present	Matters discussed
13-3	Online using Whatsapp	Group members	Discuss Requirements and search for software
143 at 8pm-9pm	Online using Whatsapp	Group members	Download various software and trial them, discuss
15-3	Online using Whatsapp	Group members	Cleaning the data and application on the software
16-3 at 6pm-8pm	Zoom	Group members	Exchange information and ideas Continue application on the software
17-3 at 8pm	Online using Whatsapp	Group members	Prepare for Presentation
27-3	Zoom	Group members	Discuss and continue application on the software
28-3	Online using Whatsapp	Group members	Discuss and continue application on the software
31-3 at 9am-1pm	King Saud University	Group members	Writing the Document
1-4 at 10am-1pm	Online using Whatsapp	Group members	Final preparation



References

-. (2008, Oct 31). *Data Mining Steps* . Retrieved from data mining warehousing: http://dataminingwarehousing.blogspot.com/2008/10/data-mining-steps-of-data-mining.html

Alton, L. (2017, Dec 22). *The 7 Most Important Data Mining Techniques*. Retrieved from Data Science Central : https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

Orange. (2019). *Orange Featuers* . Retrieved from Oreange : https://orange.biolab.si/#Orange-Features

Orange. (n.d.). *Orange Channel* . Retrieved from YouTube: https://www.youtube.com/channel/UClKKWBe2SCAEyv7ZNGhIe4g