ES 2.1 - CHAPTER 2, EXERCISE 7 PART a, b AND c

⑦ THE TABLE BELOW PROVIDES A TRAINING DATA SET CONTAINING SIX OBSERVATIONS, THREE PREDICTORS, AND ONE QUALITATIVE RESPONSE VARIABLE.

OBS.	X_1	X_2	X_3	Y
1	0	3	0	RED
2	2	0	0	RED
3	0	1	3	RED
4	0	1	2	GREEN
5	-1	0	1	GREEN
6	1	1	1	RED

SUPPOSE WE WISH TO USE THIS DATA SET TO MAKE A PREDICTION FOR Y WHEN

$X_1 = X_2 = X_3 = 0$ USING K-NEAREST NEIGHBORS

a) COMPUTE THE EUCLIDEAN DISTANCE BETWEEN EACH OBSERVATION AND THE TEST POINT, $X_1 = X_2 = X_3 = 0$.

b) WHAT IS OUR PREDICTION WITH $K = 1$? WHY?

c) WHAT IS OUR PREDICTION WITH $K = 3$? WHY?

a) DISTANCES

OBS 1: $d_1 = \sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2} = \sqrt{0+9+0} = 3$

OBS 2: $d_2 = \sqrt{(2-0)^2 + (0-0)^2 + (0-0)^2} = \sqrt{4+0+0} = 2$

OBS 3: $d_3 = \sqrt{(0-0)^2 + (1-0)^2 + (3-0)^2} = \sqrt{0+1+9} \cong 3,16$

OBS 4: $d_4 = \sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = \sqrt{0+1+4} \cong 2,24$

OBS 5: $d_5 = \sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{1+0+1} \cong 1,41$

OBS 6: $d_6 = \sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} = \sqrt{1+1+1} \cong 1,73$

b) K=1

SINCE K=1, WE SELECT ONLY THE OBSERVATION WITH THE SMALLEST DISTANCE TO THE REQUESTED POINT.

FROM THE PREVIOUS ANSWER (a) WE KNOW THAT THE CLOSEST IS THE OBSERVATION NUMBER 5 (GREEN) = 1,41. SINCE IT IS THE CLOSEST NEIGHBOR, THE Y OF $X_1=X_2=X_3=0$ IS <u>GREEN</u>.

c) K=3

IN THIS CASE WE SELECT THE 3 OBSERVATIONS WITH THE SMALLEST DISTANCE TO THE REQUESTED POINT.

FROM THE PREVIOUS ANSWER (a) WE KNOW THAT THIS TOP 3 IS COMPOSED BY:
OBSERVATION 5 (GREEN), OBSERVATION 6 (RED), OBSERVATION 2 (RED).

SINCE THE MAJORITY OF THE VALUE Y IS RED, THE PREDICTED Y OF $X_1=X_2=X_3=0$ IS <u>RED</u>

OBS	d	Y	
OBS 5	1,41	GREEN	b
OBS 6	1,73	RED	c
OBS 2	2	RED	c
OBS 4	2,24	GREEN	
OBS 1	3	RED	
OBS 3	3,16	RED	

ES 2.2 - CHAPTER 3, EXERCISE 3

SUPPOSE WE HAVE A DATA SET WITH 5 PREDICTORS, X_1 = GPA, X_2 = IQ, X_3 = LEVEL (1 FOR COLLEGE AND 0 FOR HIGH SCHOOL), X_4 = INTERACTION BETWEEN GPA AND IQ, X_5 = INTERACTION BETWEEN GPA AND LEVEL. THE RESPONSE IS STARTING A SALARY AFTER GRADUATION (IN THOUSANDS OF DOLLARS). SUPPOSE WE USE LEAST SQUARES TO FIT THE MODEL, AND GET $\hat{\beta_0} = 50$, $\hat{\beta_1} = 20$, $\hat{\beta_2} = 0,07$, $\hat{\beta_3} = 35$, $\hat{\beta_4} = 0,01$ AND $\hat{\beta_5} = -10$.

2) WHICH ANSWER IS CORRECT AND WHY?

i) FOR A FIXED VALUE OF IQ AND GPA, HIGH SCHOOL GRADUATES EARN MORE, ON AVG, THAN COLLEGE ONES

ii) // // // // //, COLLEGE GRADUATES EARN MORE, ON AVG, THAN HIGH SCHOOL ONES.

iii) // // // // //, HIGH SCHOOL GRADUATES EARN MORE, ON AVG, THAN COLLEGE ONES PROVIDED THAT THE GPA IS HIGH ENOUGH.

iv) // // // // //, COLLEGE GRADUATES EARN MORE, ON AVG, THAN HIGH SCHOOL ONES PROVIDED THAT THE GPA IS HIGH ENOUGH.

THE SALARY PREDICTION MODEL IS: $Y = 50 + 20X_1 + 0,07X_2 + 35X_3 + 0,01X_4 - 10X_5$

IN PARTICULAR, FOR HS GRADUATES: $Y_{HS} = 50 + 20X_1 + 0,07X_2 + 35 \cdot 0 + 0,01X_4 - 10X_1 \cdot 0 =$
$= 50 + 20X_1 + 0,07X_2 + 0,01X_4$

AND FOR COLLEGE GRADUATES: $Y_{COL} = 50 + 20X_1 + 0,07X_2 + 35 \cdot 1 + 0,01X_4 - 10X_1 =$
$= 85 + 10X_1 + 0,07X_2 + 0,01X_4$

SO, THE SALARY DIFFERENCE BETWEEN HS AND COL GRADUATES IS GIVEN BY THIS COMPARISON:

$Y_{COL} - Y_{HS} = (85 + 10X_1 + 0,07X_2 + 0,01X_4) - (50 + 20X_1 + 0,07X_2 + 0,01X_4) =$
$= 35 - 10X_1$ $\rightarrow X_1 = \frac{35}{10} = 3,5$ IS THE THRESHOLD

2 OPTIONS:

1) IF X_1 = GPA IS LOW, $35 - 10X_1$ IS POSITIVE, SO COL GRADUATES EARN MORE

2) IF X_1 = GPA IS HIGH, $35 - 10X_1$ IS NEGATIVE, SO HS GRADUATES EARN MORE

SO, SINCE HS GRADUATES EARN MORE THAN COL ONES ONLY IF GPA IS HIGH ENOUGH (MORE THAN 3,5), THE CORRECT ANSWER IS THE NUMBER iii.

b) PREDICT THE SALARY OF A COLLEGE STUDENT WITH IQ OF 110 AND A GPA OF 4,0

IN THIS CASE I USE THE PREVIOUS Y_{COL} MODEL KNOWING THAT:

$X_1 = 4,0 \quad X_2 = 110 \quad$ AND $\quad X_3 = 1$

SO, $X_4 = X_1 \cdot X_2 = 4 \cdot 110 = 440 \quad$ AND $\quad X_5 = X_1 \cdot X_3 = 4 \cdot 1 = 4$

AND SO,

$$Y_{COL} = 50 + 20 \cdot 4 + 0,07 \cdot 110 + 35 \cdot 1 + 0,01 \cdot 440 - 10 \cdot 4 =$$
$$= 50 + 80 + 7,7 + 35 + 4,4 - 40 =$$
$$= 137,1$$

WHICH MEANS THAT THIS COLLEGE STUDENT WILL EARN A SALARY EQUAL TO 137'100 DOLLARS

c) TRUE OR FALSE: SINCE THE COEFFICIENT FOR THE GPA - IQ INTERACTION TERM IS VERY SMALL, THERE IS VERY LITTLE EVIDENCE OF AN INTERACTION EFFECT. JUSTIFY YOUR ANSWER.

THIS IS FALSE, BECAUSE EVEN IF THE $\beta_4 = 0,01$ IS VERY SMALL, THE X_4 CAN HAVE HIGH VALUES, AND AS WE SAW IN THE ANSWER (b) THE END RESULT OF THIS INTERACTION IS STILL RELEVANT. TO UNDERSTAND BETTER IF THIS INTERACTION HAS AN EFFECT ON THE MODEL WE SHOULD STUDY ITS SIGNIFICANCE WITH THE P-VALUE (P-VALUE $< 0,05$ = SIGNIFICANT).

ES 2.3 – CHAPTER 4, EXERCISE 6

SUPPOSE WE COLLECT DATA FROM A GROUP OF STUDENTS IN A STATISTICS CLASS, WITH VARIABLES X_1 = HOURS STUDIED, X_2 = UNDERGRAD GPA, AND Y = RECEIVE AN A.

WE FIT A LOGISTIC REGRESSION AND PRODUCE ESTIMATED COEFFICIENT, $\beta_0 = -6$, $\beta_1 = 0{,}05$, $\beta_2 = 1$.

a) ESTIMATE THE PROBABILITY THAT A STUDENT WHO STUDIES FOR 40 h AND HAS AN UNDERGRAD GPA = 3,5 GETS AN A IN THE CLASS.

FIRSTLY, THE LOGISTIC REGRESSION MODEL, GIVEN THE COEFFICIENTS, IS:

$$P(Y=1) = \frac{e^{-6+0{,}05X_1+1X_2}}{1+e^{-6+0{,}05X_1+1X_2}}$$

AND KNOWING THAT $X_1 = 40$ AND $X_2 = 3{,}5$ IT BECOMES:

$$P(Y=1) = \frac{e^{-6+0{,}05\cdot 40+1\cdot 3{,}5}}{1+e^{-6+0{,}05\cdot 40+1\cdot 3{,}5}} = \frac{e^{-6+2+3{,}5}}{1+e^{-6+2+3{,}5}} = \frac{e^{-0{,}5}}{1+e^{-0{,}5}} \cong$$

$$\cong \frac{0{,}606}{1+0{,}606} = 0{,}377 \rightarrow \boxed{P = 37{,}7\%}$$

b) HOW MANY HOURS WOULD THE STUDENT IN PART (a) NEED TO STUDY TO HAVE A 50% CHANCE OF GETTING AN A IN THE CLASS?

IN THIS CASE WE NEED TO FIND THE X_1 S.T. $P(Y=1) = 0{,}5$.

SO,

$$\frac{e^{-6+0{,}05X_1+1\cdot 3{,}5}}{1+e^{-6+0{,}05X_1+1\cdot 3{,}5}} = 0{,}5 \rightarrow e^{(\dots)} = 0{,}5\left(1+e^{(\dots)}\right) \rightarrow$$

$$\rightarrow e^{-6+0{,}05X_1+3{,}5} - 0{,}5\cdot e^{-6+0{,}05X_1+3{,}5} = 0{,}5$$

$$\rightarrow 0{,}5\cdot e^{-6+0{,}05X_1+3{,}5} = 0{,}5^1 \rightarrow \ln\left(e^{(\dots)}\right) = \ln(1) = 0$$

$$\rightarrow -6+0{,}05\cdot X_1+3{,}5 = 0 \rightarrow 0{,}05X_1 = 2{,}5 \rightarrow X_1 = \frac{2{,}5}{0{,}05} = 50$$

THE STUDENT WILL NEED TO STUDY FOR 50 HOURS.

• Es 2.4 - Topic DM

A retailer wants for marketing purposes distinguish btwn costumers younger than 35 and older than 35. The following table summarizes the data stat in the data base of the retailer in an abstract form. The relevant attributes, determined by domain knowledge, are for convenience denoted by A and B. The values for A are a_1, a_2 and a_3. The values for B are b_1 and b_2.

The retailer wants to use the DM techniques to classify the customers in the class "Young", denoted by "Y", and "old" denoted by "O".

A	B	# of instances Y	O
a_1	b_1	4	10
a_2	b_1	6	2
a_3	b_1	8	6
a_1	b_2	2	8
a_2	b_2	6	2
		Tot: 26	Tot: 28

a) Assume a new costumer enters the web-store and the retailer has no information at all about this customer. How will this new costumer be classified based on the above data and explain why.

In this case, since no information about the age of the costumer is given, the retailer should try to predict if his class is "Y" or "O" using the data in the table in this way:

Total Y = 4 + 6 + 8 + 2 + 6 = 26

Total O = 10 + 2 + 6 + 8 + 2 = 28

Since "O" has more instances, the probable prediction for the new costumer is "O".

b) Now assume a new costumer comes in for which the retailer knows the values for attribute $A = a_3$ and $B = b_2$. Is it possible to apply the standard (naive Bayes) to classify this new customer? Explain what the problem is.

In this case, to apply the Standard Bayes we need to calculate the probability of the customer belonging to both "Y" and "O" knowing that $A = a_3$ and $B = b_2$.

Rule of Bayes:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

So:

$$P(Y|a_3,b_2) = \frac{P(a_3,b_2|Y)\cdot P(Y)}{P(a_3,b_2)} \quad ; \quad P(0|a_3,b_2) = \frac{P(a_3,b_2|0)\cdot P(0)}{P(a_3,b_2)}$$

With: $P(a_3,b_2|Y)$, $P(a_3,b_2|0)$ and $P(a_3,b_2)$ unknown.

Unfortunately in this case we cannot apply the Simple Bayes because in the table there is no data with a combination of a_3 and b_2 in the same row, infact $P(A=a_3, B=b_2) = 0$.

c) Have the lecturer decides to use a Naive Bayes classifier for the classification of this new costumer. How will this customer now be classified based on the values of A and B? Explain your answer.

Using the same probability notations of the answer (b), and knowing that the denominator $P(a_3,b_2)$ is the same for both $P(Y|a_3,b_2)$ and $P(0|a_3,b_2)$ we can ignore them (but following the "Matlab Lecture" slides, we computed it).

So, going on, we can say that a_3 and b_2 are conditional independent which means:

$$P(a_3,b_2|Y) = \underbrace{P(a_3|Y)}_{\text{known from table}}\cdot\underbrace{P(b_2|Y)}_{\text{known from table}} = \frac{8}{26}\cdot\frac{2+6}{26} = \frac{\cancel{8}^{4}}{\cancel{26}_{13}}\cdot\frac{8^{4}}{\cancel{26}_{13}} = \frac{16}{169} = 0{,}0946$$

$$P(a_3,b_2|0) = \underbrace{P(a_3|0)}_{\text{known from table}}\cdot\underbrace{P(b_2|0)}_{\text{known from table}} = \frac{6}{28}\cdot\frac{3+2}{28} = \frac{\cancel{6}^{3}}{\cancel{28}_{14}}\cdot\frac{\cancel{10}^{5}}{\cancel{28}_{14}} = \frac{15}{196} = 0{,}0765$$

Also

Denominator

$$\begin{aligned} P(a_3,b_2) &= P(a_3,b_2|Y)\cdot P(Y) + P(a_3,b_2|0)\cdot P(0) = \\ &= P(a_3|Y)\cdot P(b_2|Y)\cdot P(Y) + P(a_3|0)\cdot P(b_2|0)\cdot P(0) = \\ &= 0{,}0946\cdot\frac{\cancel{26}^{13}}{\cancel{54}_{27}} + 0{,}0765\cdot\frac{\cancel{28}^{14}}{\cancel{54}_{27}} = 0{,}0455 + 0{,}0396 = 0{,}0851 \end{aligned}$$

So,

$$P(Y|a_3,b_2) = \frac{\cancel{2}^{1}\cancel{6}^{13}}{\cancel{26}_{13}}\cdot\frac{8}{\cancel{26}}\cdot\frac{\cancel{26}}{\cancel{54}_{27}} = \boxed{\frac{16}{351}}$$

$$P(0|a_3,b_2) = \frac{\cancel{3}^{1}\cancel{6}^{3}}{\cancel{28}_{14}}\cdot\frac{\cancel{5}^{1}\cancel{10}^{5}}{\cancel{28}_{14}}\cdot\frac{\cancel{28}^{14}}{\cancel{54}_{27}} = \boxed{\frac{15}{378}}$$

Since $\frac{16}{351} > \frac{15}{378}$

the probability of "Y" is higher, so the new customer is classified as "Y"

• ES 2.5 - TOPIC DM

CONSIDER THE SAME DATASET AS IN THE PREVIOUS QUESTION. NOW THE RESEARCH (DM ANALYST) WANTS TO USE DECISION TREES TO CLASSIFY NEW CUSTOMERS.

a) WHAT IS THE CLASSIFICATION ERROR RATE FOR ATTRIBUTE A?

A	Y	O	LABEL	# OF ERRORS
a_1	6	18	O	6
a_2	12	4	Y	4
a_3	8	6	Y	6

$$\text{CLASS ERROR RATE}(A) = \frac{6+4+6}{54} = \frac{16}{54} = 0,296$$

b) WHAT IS THE CLASSIFICATION ERROR RATE FOR ATTRIBUTE B?

B	Y	O	LABEL	# OF ERRORS
b_1	18	18		18
b_2	8	10		8

$$\text{CLASS ERROR RATE}(B) = \frac{18+8}{54} = \frac{26}{54} = 0,481$$

c) WHAT WILL BE THE SPLITTING ATTRIBUTE IN THE TOP (ROOT) OF THE DECISION TREE IF WE USE THE CLASSIFICATION ERROR RATE?

SINCE THE CHOICE IS BASED ON THE ATTRIBUTE WITH THE LOWEST ERROR RATE, THE ROOT OF THE DECISION TREE WILL BE A $(0,296 < 0,481)$.

d) WHAT IS THE GINI INDEX FOR ATTRIBUTE A?

GINI INDEX FOR ATTRIBUTE A: $GI(\text{NODE } M) = \sum_{k=1}^{K} \hat{p}_{Mk}(1-\hat{p}_{Mk})$, WHERE $\hat{p}_{Mk}$ IS THE PROPORTION OF CLASS k OBSERVATIONS IN NODE M.

$$GI(a_1) = \frac{6}{6+18}\cdot\frac{18}{6+18} + \frac{18}{6+18}\cdot\frac{6}{6+18} = \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{9}{12} = \frac{27}{72} = 0,375$$

$$GI(a_2) = \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{4} = \frac{3}{8} = 0,375$$

$$GI(a_3) = 2\cdot\frac{4}{7}\cdot\frac{3}{7} = \frac{24}{49} = 0,490$$

$$\text{OVERALL } GI(A) = \frac{24}{54}\cdot 0,375 + \frac{16}{54}\cdot 0,375 + \frac{14}{54}\cdot 0,490 = 0,4048$$

e) WHAT IS THE GINI INDEX FOR ATTRIBUTE B?

$$GI(b_1) = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = 0,5$$

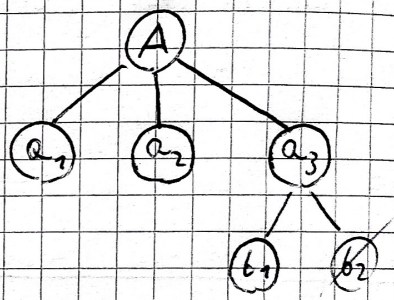
$$GI(b_2) = 2 \cdot \frac{4}{9} \cdot \frac{5}{9} = \frac{40}{81} = 0,4938$$

OVERALL $GI(B) = \frac{2}{3} \cdot 0,5 + \frac{1}{3} \cdot 0,4938 = \frac{2}{3} \cdot 0,5 + \frac{1}{3} \cdot 0,4938 = 0,498$

f) WHAT WILL BE THE SPLITTING ATTRIBUTE IN THE TOP (ROOT) OF THE DECISION TREE IF ONE USES THE GINI INDEX?

SINCE THE CHOICE IS BASED ON THE ATTRIBUTE WITH THE LOWEST ERROR RATE, THE ROOT OF THE DECISION TREE WILL BE A ($0,4048 < 0,498$).

g) CONSTRUCT THE FULL DECISION TREE, USING THE ERROR RATE AS MEASURE, AND WHAT IS THE OVERALL CLASSIFICATION ERROR RATE ON THE ABOVE DATASET?



FROM ANSWERS (a) AND (b) THE ROOT IS (A)

THEN WE SPLIT A INTO ITS 3 BRANCHES: a_1, a_2, a_3

ERR RATE $(a_1) = \frac{6}{24} = 0,25$

ERR RATE $(a_2) = \frac{4}{16} = 0,25$

ERR RATE $(a_3) = \frac{6}{14} = 0,43$

SINCE THIS $a_3 = 0,43$ IS SLIGHTLY HIGHER THAN THE PREVIOUS TWO, WE DECIDE THAT $a_1 \rightarrow O$, $a_2 \rightarrow Y$ AND $a_3 \rightarrow$ SPLIT INTO b_1 AND b_2 (BUT WE HAVE ONLY b_1).

ERR RATE $(a_3, b_1) = \frac{6}{14} = 0,43$ SO WE CHOOSE THIS GROUP AS $Y (8 > 6)$.

OVERALL CLASSIFICATION ERR RATE $= \frac{6+4}{54} = 0,203$

h) IS THIS CLASSIFICATION ERR RATE AN OPTIMISTIC OR PESSIMISTIC ESTIMATE OF THE ERR RATE ON UNSEEN NEW DATA? EXPLAIN YOUR ANSWER

SINCE THE ACCURACY IS $1 - 0,203 = 0,797$, WE CAN SAY THAT IT IS AN OPTIMISTIC ESTIMATE

* TOPIC DEP

PAPAR AND PEN EXERCISES: ES 1.3.2 - ASSIGNMENT 1: FACTS AND DIMENSIONS

ES 1.3.5 - ASSIGNMENT 4: "MOBILE APP B-TREE SERVICE"

PROGRAMMING EXERCISES: ES 1.3.3 - ASSIGNMENT 2: A SIMPLE ETL PROGRAM TO START WITH

ES 1.3.4 - ASSIGNMENT 3: DO IT YOURSELF

ES 1.3.2 - ASSIGNMENT 1: FACTS AND DIMENSIONS

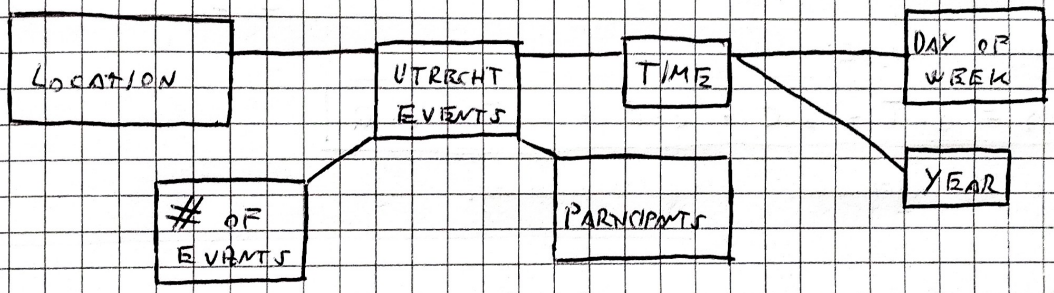
SEE FIGURE 1.1 WHICH CONTAINS DATA ON EVENTS HAPPENING IN THE CITY OF UTRECHT. WE ILLUSTRATE THE MEANING OF THE DATA BY EXPLAINING WHAT THE FIRST ROW MEANS: THERE WERE IN TOTAL 2 EVENTS ORGANISED IN EUCLIDEAN ON A FRIDAY IN 2016 WHICH BOTH TOGETHER DREW 15'000 PARTICIPANTS.

a) WHICH ATTRIBUTES ARE THE FACTS; WHICH ATTRIBUTES ARE THE DIMENSIONS?

FACTS (NUMERIC DATA): "#ofevents" AND "totalparticipants"

DIMENSIONS (DESCRIPTIVE DATA): "dayofweek", "year", AND "location".

b) DRAW A CONCEPTUAL STAR SCHEMA FOR THIS CUBE

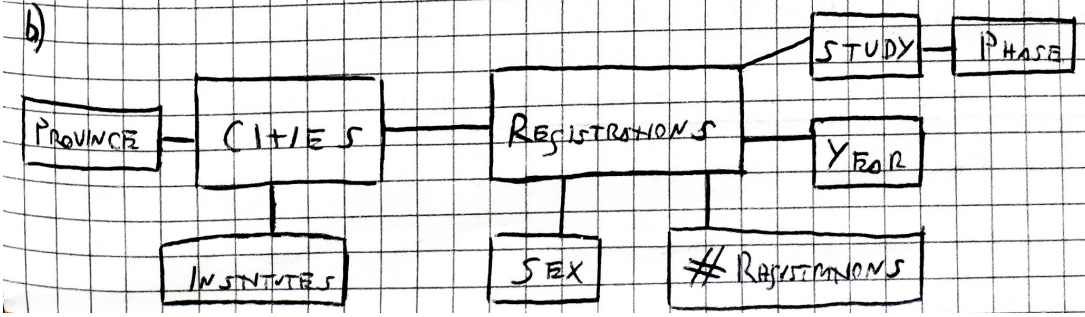


SEE FIGURE 1.2 WHICH CONTAINS DATA ON THE NUMBERS OF REGISTERED STUDENTS IN THE NETHERLANDS. WE ILLUSTRATE THE MEANING OF THE DATA BY EXPLAINING WHAT THE FIRST ROW OF TABLE "REGISTRATIONS" MEANS: IN 2015 THERE WERE 13 FEMALE STUDENTS REGISTERED FOR THE BACHELOR STUDY "TECHNICAL COMPUTER SCIENCE" AT THE UT WHICH IS LOCATED IN THE CITY OF ENSCHEDE IN THE PROVINCE OF OVERIJSSEL.

a) FACTS (NUMERIC DATA): "#ofregistrations"

DIMENSIONS (DESCRIPTIVE DATA): "instID", "study", "phase", "year" AND "sex" (AND "cityID").

b)



ES 1.3.5 - Assignment 4: Multidimensional modeling. Case "Mobile App β testing service"

a) For being able to determine the right fact(s) for the Star Schema an important question is "What is good", i.e., "What makes a particular game a good β tester for a particular kind of game?". More concretely, how can you determine a score that summarizes this "goodness" or "suitability", which obviously needs to be calculated from the available data. Propose a formula for "goodness score" and explain why a high value is in favor for a good β tester.

To answer the question "What makes a particular game a good β tester for a particular kind of game?" we thought about these characteristics/requirements:

1) Plays the game for a good amount of time (T) [time spent] → the higher, the better
2) Plays consistently during the testing period (C) [consistence] → " " , " "
3) Is he/she good at playing it? (P) [Pro-level] → " " , " "
4) Provides meaningful feedbacks? (F) [Feedback] → " " , " "
5) # of matches/duels/levels quitted in the testing period. (Q) [Quitting rate] → the higher, the worse

So, a possible formula could be:

$$\boxed{G = a \cdot T + b \cdot C + c \cdot P + d \cdot F - e \cdot Q}$$

with a, b, c, d, e as coefficients of the requirements.

b) What is the business question, or what are the business questions in this case? Formulate them as accurately as possible.

1) How can Pear find the perfect β-tester for a specific game?
   → Should Pear choose them between streamers/youtubers, or also between common people who enjoy the game's genre? Only 1 tester or more? From which country/countries? How can Pear know that the tester's feedbacks are trustworthy?
2) How can Pear measure the effectiveness of its service?
   → Which data can Pear collect during this process? Should Pear share all the data to the developers? How to deal with the privacy of the testers? Which KPIs should Pear use?

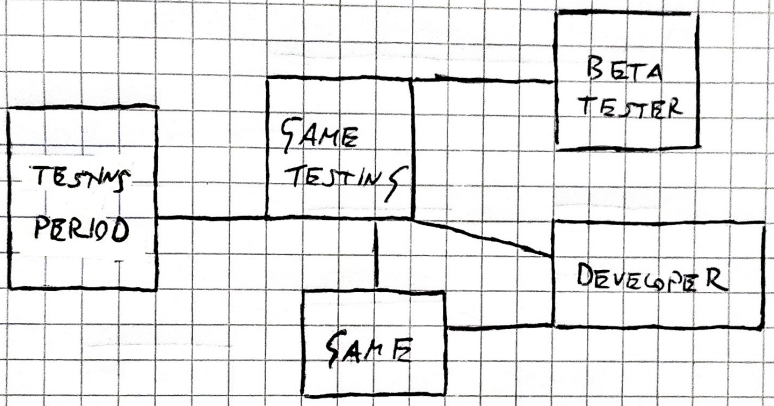
3) How can PBar monetize its service?

-> Which is the best pricing strategy? How much should the developers pay for the subscription, and how much for hiring the tester? More testers = more expenses? or maybe a "tester group" discounted price? Will loyal customers (developers) receive discounts in subsequent years?

4) How can PBar optimize/improve its service?

-> Which data can be used to find better tester-developer matches? Should PBar take care of developers feedbacks before/after the service? Should PBar use external comments to augment the server level, or an internal team is enough?

c) Give a Star Schema. Explain your design by listing the most important design choices and considerations



The center of the Star Schema is the "Game Testing" table which contains the data and the info of a Testing session (Test ID = Primary Key). It is connected to 4 other tables through 4 Foreign Keys: BTesterID, DeveloperID, GameID, PeriodID.

The "Beta Tester" table contains all the info about the β tester (e.g: ID, Name, Age, Genre, Feedbacks rate...).

The "Developer" table contains all the info about the Developer (e.g: ID, company name, # of games released...).

The "Game" table contains all the info about the Game (e.g: ID, Genre, # of levels, multiplayer, DeveloperID [FK]...).

The "Testing Period" table contains all the info about the Time (e.g: ID, starting date, ending date, year...).