CHALMERS

EXAMINATION / TENTAMEN

Course code/kurskod	Course name/kursnamn		
DIT852	Introduction to Data Science		
Anonymous code Anonym kod	Examination date Tentamensdatum	Number of pages Antal blad	Grade Betyg
DIT852-0001-ZCG	Oct/24/2023	10	

I confirm that I've no mobile or other similar electronic equipment available during the examination.
Jag intygar att jag inte har mobiltelefon eller annan liknande elektronisk utrustning tillgänglig under eximinationen.

Solved task Behandlade uppgifter No/nr		Points per task Poäng på uppgiften	Observe: Areas with bold contour are to completed by the teacher. Anmärkning: Rutor inom bred kontur ifylles av lärare.
1	✓	6	
2	✓	4.5	
3	✓	6	
4	✓	6	
5	✓	5	
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
Bonus poäng	3		
Total examination points Summa poäng på tentamen		30,5	

[redacted]

CHALMERS	Anonymous code	Points for question (to be filled in by teacher)	Consecutive page no. Löpande sid nr 1
	Anonym kod DIT852-0001-ZCG	Poäng på uppgiften (ifylles av lärare)	Question no. Uppgift nr 1

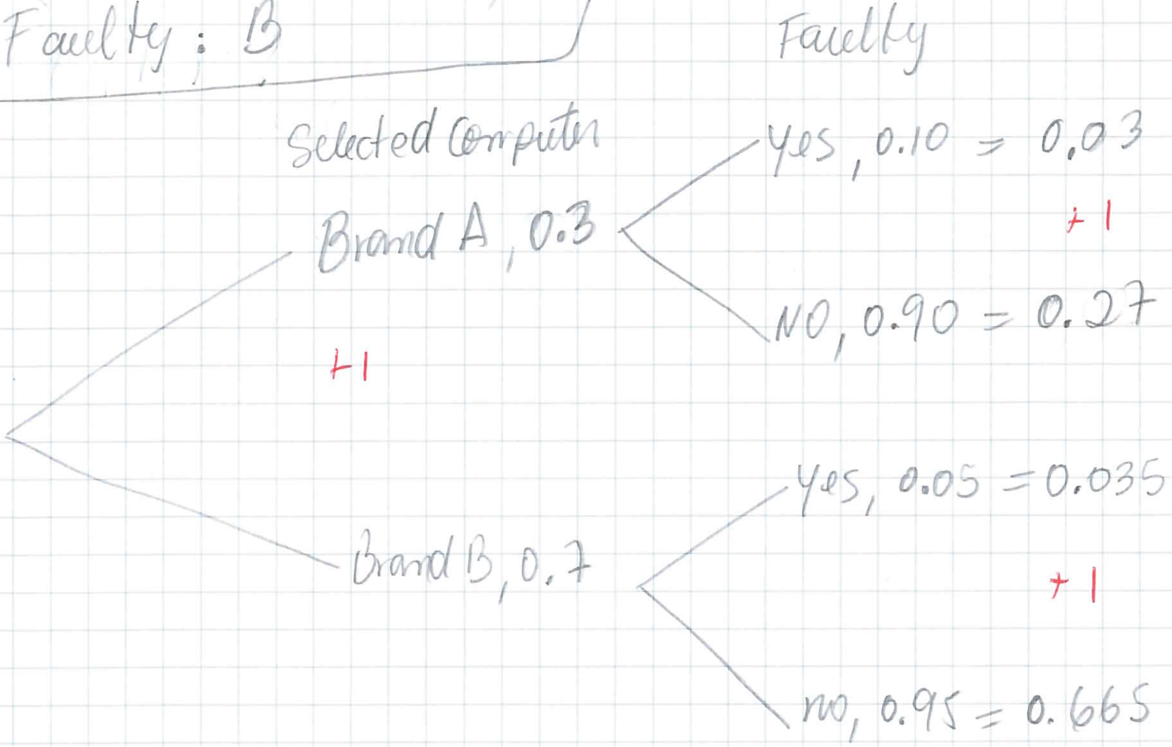
Question 1

Brand A: OCP: 10%
Brand B. Cyberdine: 5% → faulty +1

30% of the computers sold are manufactured by Brand A.

What is the probability that a random selected faulty computer is OCP?

Selected Computer OCP: A
Faulty: B
→ $P(A|B) = ?$



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.03}{0.03 + 0.035} = 0.46 //$$

+1 +1

The probability that a random selected faulty computer is OCP is 0.46

CHALMERS	Anonymous code Anonym kod DIT852-0001-ZCG	Points for question (to be filled in by teacher) Poäng på uppgiften (ifylles av lärare)	Consecutive page no. Löpande sid nr 2 Question no. Uppgift nr 2

Question 2

a) They are stored as strings, which make difficult because we need to transform them to datetime.

The challenges are that:

i) Countries follow different formats month/day/year others day/month/year. +1

ii) Regarding time some parts of the world use 24 hours others 12 hours (9.a.m, 9 p.m)

iii) Sometimes they input the date the way they are used to e.g. Oct 24th (US) Oct 24 (UK)

+1

b) Because time differs depending on the time zone. For example, in New Year Eve, 1 of January starts first in Asia and Australia and then to EUROPE and then to AMERICA. ×

I would separate the dataset between time zones, so I can understand what exactly what is happening in each time zone.

Another way, should be format or transform all dates to the timezone where I'm located, +1 but I will be careful with my conclusions, keeping in mind that there are information from other time zones.

CHALMERS	Anonymous code	Points for question (to be filled in by teacher)	Consecutive page no. Löpande sid nr 3
	Anonym kod DIT852-0001-ZCG	Poäng på uppgiften (ifylles av lärare)	Question no. Uppgift nr 2

Question 2

c) i) February is always 28 days long. False, every 4 years february has 29 days (adjustment for our calendar) +0,5

ii) There are 24 hours in a day. False, some countries use 12 hours (10 a.m, 10 p.m) +0,25

iii) The offsets between two time zones will remain constant. False, the offset between Sweden and Ecuador is 7 hours when is spring and summer in Sweden, then is 6 hours in late autumn and winter. +0,5

iv) There is a leap year every year divisible by four. False, leap year are not base whether a year is divisible by four. We can say that after four year, we have a leap year. +0.25

2

CHALMERS	Anonymous code Anonym kod DIT852-0001-ZCG	Points for question (to be filled in by teacher) Poäng på uppgiften (ifylles av lärare)	Consecutive page no. Löpande sid nr 4 Question no. Uppgift nr 3

Question 3

a) KNN use labeled data to perform classification
Second, we select K, K corresponds to the number of nearest neighbors or closet points to a new unseen point (Closet points based on a distance metric such as: Euclidean, Manhattan, Angular)
Third, we select our distance metric (usually Euclidean), then K-NN will calculate the distance between the unseen point and the K nearest neighbors, these distances are ranked from smallest to largest. Finally, KNN will assign the most repetitive label of K values ~~to~~ that unseen point. e.g.

2

K = 5   categories = blue, red   q = query point)

$d_1 = D(q, K_1)$: ⟷ K_1 = blue
$d_2 = D(q, K_2)$: ⟷ K_2: red
$d_3 = D(q, K_3)$: ⟷ K_3: blue
$d_4 = D(q, K_4)$: ⟷ K_4: blue
$d_5 = D(q\ K_5)$: ⟷ K_5: red

blue will be the label for q.

CHALMERS	Anonymous code Anonym kod DIT852-0001-ZCG	Points for question (to be filled in by teacher) Poäng på uppgiften (ifylles av lärare)	Consecutive page no. Löpande sid nr 5 Question no. Uppgift nr 3

Question 3

b) The curse of dimensionaly refers to the challenges that we face when we are dealing with high-dimension data, data set with a high numbers of variables.

Some of them:

i) Data Sparsity: in high dimension data point tend to be more spread out (scaterred) making difficult to find relationship between them.

ii) High Compuetational Needs: to analyze data in high dimension we need more computational power, and this is not always possible. 2

iii) Diminished Intituition: It becomes harder for us to analyze or visualize data in high dimensions, difficulting getting insight from the data.

iv) Thereare some techniques that we can use to reduce dimensionality such as PCA, Random Projections. However, we will be working with group rather than the features.

KNN use the distance metric of K points to assign a label to an unseen point. in high-dimensions it will really difficult for KNN to calculate those distances because of the data sparsitey and the high computational power needed.

CHALMERS	Anonymous code Anonym kod DIT852-0001-ZCG	Points for question (to be filled in by teacher) Poäng på uppgiften (ifylles av lärare) 6	Consecutive page no. Löpande sid nr 6 Question no. Uppgift nr 3

Question 3

c) - K = the number of nearest neighbor considered to assign a label to an unseen data point based on the distema metric

- a small K, does not provide a robust classification
- a high K, makes K-NN less robust, specially in cases when K=n because the algorithm will just assign an unseen point to the majority label (dummy classifier).

2

- to choose K, we can plot the accuracy on our testing dataset for different values of K. Check the grahp an identified for which K we have kane highest accuracy, that should be the best K.

2

CHALMERS	Anonymous code Anonym kod DIT852-0001-ZCG	Points for question (to be filled in by teacher) Poäng på uppgiften (ifylles av lärare)	Consecutive page no. Löpande sid nr 7 Question no. Uppgift nr 4

Question 4

a)

PEARSON	SPEARMAN
• it quantifies the linear relationship betwem two continous variables	• it assess the monotonic relationship between two continuous variables or ordinal (ranked) variables
• Sensitive to outliers	• It is not sensitive to outliers because it works with ranked values rather than the data points.
• It considers the covariance and standardize by the product of the standard deviation.	• It uses the pearson correlation on the ranked values
• It's better to use it on linear relationship like weight and height.	• It's better to use for non-linear relationship in variables. e.g. grades

2

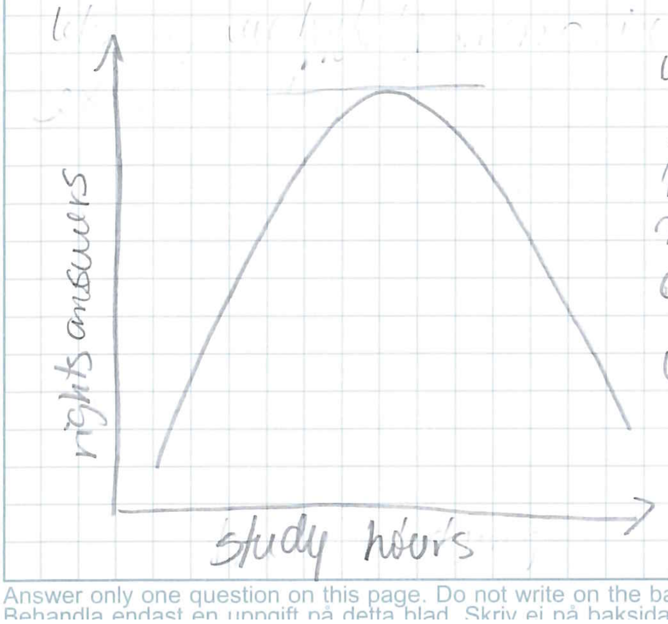
Answer only one question on this page. Do not write on the back of this paper
Behandla endast en uppgift på detta blad. Skriv ej på baksidan

Question 4

b) By visual inspection, there is a positive correlation between ice cream sales and deaths by drowning in Sweden. We can see that both variables move on the same direction in most of the months, the r could be between 0.5-1 (which is high positive). This indicates that as the sales of ice cream in Sweden increases, the deaths by drowning also increases. Also that, as the sales of ice cream decreases, the same is for the death by drowning. However, correlation does not mean causation. But also, with some intuition, we can say that we can get the correlation between two random varibles, but also we should ask ourselves wether this make sense (ice cream vs death by drowning). 2

c) It means that there is no LINEAR relationship, but there may be possible that the two variables are related in a NON-LINEAR way. 2

e.g hours spent on studying vs right answers on questionaire



we could say that the relationship between study hours and right answers on my questionaire are positive correlated, but that's true until certain points. Because after spending 20 hours studying one will feel very tired and the number of correct answers decreases.

Answer only one question on this page. Do not write on the back of this paper
Behandla endast en uppgift på detta blad. Skriv ej på baksidan.

Question 5   F A T

Fairness:

It is clear that the training data was biased. There was a group with no enought representation to train the model, which make the algorithm to incorrectly labeled them as gorillas. But, maybe there were other subpopulation that were not weighted accordingly which make this an unfair process.

Accountability:

there was accountability by recognizing the mistake, but no in how they deploy their model, they didn't perform all the neccesary checks to make sure this kind of problems do not happen. This also suggest that the testing data was also imbalanced.

Transparency:

Google was not transparent about what data the used, whether the data was imbalanced, and wether this repeats on the testing data. they also do not explain of the algorithm works or learns when people correct mistakes on the labels for photos.

2

CHALMERS	Anonymous code Anonym kod 011852-0001-2CG	Points for question (to be filled in by teacher) Poäng på uppgiften (ifylles av lärare) 5	Consecutive page no. Löpande sid nr 10 Question no. Uppgift nr 5

Question 5

This is the result of not implementing ethical practices such as balanced dataset, performance on minority groups.

One way to prevent this is by assigning weight offset the imbalance, also they could first check how many observations they have for each person race and try to make sure they gather more data if it was necessary. Also make sure that the test data has an enough number of observations for each race.

The decision of removing the tag of gorillas was correct, but also they should check whether this is happening with other labels.

Answer only one question on this page. Do not write on the back of this paper
Behandla endast en uppgift på detta blad. Skriv ej på baksidan