CHƯƠNG 1: CÂU HỎI LÝ THUYẾT

Câu 1) Present discrete attribute and continuous attribute. Take examples to show the differences.

Discrete Attribute

- Has only a finite or countably infinite set of values
  Ex1: Zip codes, profession, or the set of words in a collection of documents
  Ex2: The number of students in a class we can't have half a student
  Ex3: The results of rolling 2 dice Only have the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 we can not have 2.1 or 3.5
- Sometimes, represented as integer variables
- Binary attributes are a special case of discrete attributes

Continuous Attribute

- Has real numbers as attribute values
  Ex: Temperature, height, weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuos attributes are typically represented as floating-point variables

Câu 2) Present advantages and disadvantages of Decision Tree.

Advantages:

- It can be used for both classification and regression problems: Decision trees can be used to predict both continuous and discrete values i.e. they work well in both regression and classification tasks.
- As decision trees are simple hence they require less effort for understanding an algorithm.
- It can capture nonlinear relationships: They can be used to classify non-linearly separable data.
- An advantage of the decision tree algorithm is that it does not require any transformation of the features if we are dealing with non-linear data because decision trees do not take multiple weighted combinations into account simultaneously.
- They are very fast and efficient compared to KNN and other classification algorithms.
- Easy to understand, interpret, visualize.

- The data type of decision tree can handle any type of data whether it is numerical or categorical, or boolean.
- Normalization is not required in the Decision Tree.
- The decision tree is one of the machine learning algorithms where we don’t worry about its feature scaling. Another one is random forests. Those algorithms are scale-invariant.
- It gives us and a good idea about the relative importance of attributes.
- Useful in data exploration: A decision tree is one of the fastest way to identify the most significant variables and relations between two or more variables. Decision trees have better power by which we can create new variables/features for the result variable.
- Less data preparation needed: In the decision tree, there is no effect by the outsider or missing data in the node of the tree, that’s why the decision tree requires fewer data.
- Decision tree is non-parametric: Non-Parametric method is defined as the method in which there are no assumptions about the spatial distribution and the classifier structure.

Disadvantages:

- Concerning the decision tree split for numerical variables millions of records: The time complexity right for operating this operation is very huge keep on increasing as the number of records gets increased decision tree with to numerical variables takes a lot of time for training.
- Similarly, this happens in techniques like random forests, XGBoost.
- Decision tree for many features: Take more time for training-time complexity to increase as the input increases.
- Growing with the tree from the training set: Overfit pruning (pre, post), ensemble method random forest.
- Method of overfitting: If we discuss overfitting, it is one of the most difficult methods for decision tree models. The overfitting problem can be solved by setting constraints on the parameters model and pruning method.
- As you know, a decision tree generally needs overfitting of data. In the overfitting problem, there is a very high variance in output which leads to many errors in the final estimation and can show highly inaccuracy in the output. Achieve zero bias (overfitting), which leads to high variance.

- Reusability in decision trees: In a decision tree there are small variations in the data that might output in a complex different tree is generated. This is known as variance in the decision tree, which can be decreased by some methods like bagging and boosting.
- It can't be used in big data: If the size of data is too big, then one single tree may grow a lot of nodes which might result in complexity and leads to overfitting.
- There is no guarantee to return the 100% efficient decision tree.

Câu 3) Present advantages and disadvantages of Support Vector Machine.

Advantages:

- Give good results even if there is not enough information about the data. Also works well with unstructured data.
- Solves complex problems with a convenient kernel solution function.
- Relatively good scaling of high – dimensional data.
- SVM classifiers perform well in high-dimensional space and have excellent accuracy. SVM classifiers require less memory because they only use a portion of the training data.
- SVM performs reasonably well when there is a large gap between classes.
- High-dimensional spaces are better suited for SVM.
- When the number of dimensions exceeds the number of samples, SVM is useful.
- SVM uses memory effectively.

Disadvantages:

- It is difficult to choose the appropriate kernel solution function.
- Training time is long when using large data sets.

- It may be difficult to interpret and understand because of problems caused by personal factors and the weights of variables.
- The weights of the varibales are not constant, thus the contribution of each variable to the output is variant.
- SVM requires a long training period; as a result, it is not practical for large datasets.
- The inability of SVM classifiers to handle overlapping classes is another drawback.
- Large data sets are not a good fit for the SVM algorithm.
- When the data set contains more noise, such as overlapping target classes, SVM does not perform as well.
- The SVM will perform poorly when the number of features for each data point is greater than the number of training data samples.

Câu 4) Give one example of data mining application in the field of education or public transportation. Based on your example, what kind of data, and data mining method you can use?

Câu 5) Compare supervised learning with unsupervised learning. Take examples to show the differences.

Supervised learning: is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.

Supervised learning can be separated into two types of problems when data mining: classification and regression:

- Classification problems use an algorithm to accurately assign test data into specific categories, such as separating apples from oranges.
- Regression is another type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables. Unsupervised learning: uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention

Unsupervised learning: uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention

Unsupervised learning models are used for three main tasks: clustering, association and dimensionality reduction:

- Clustering is a data mining technique for grouping unlabeled data based on their similarities or differences.
- Association is another type of unsupervised learning method that uses different rules to find relationships between variables in a given dataset.
- Dimensionality reduction is a learning technique used when the number of features (or dimensions) in a given dataset is too high

The main distinction between the two approaches is the use of labeled datasets. To put it simply, supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.

In supervised learning, the algorithm "learns" from the training dataset by iteratively making predictions on the data and adjusting for the correct answer.

- For example, a supervised learning model can predict how long your commute will be based on the time of day, weather conditions and so on. But first, you'll have to train it to know that rainy weather extends the driving time.

Unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabeled data.

- For example, an unsupervised learning model can identify that online shoppers often purchase groups of products at the same time. However, a data analyst would need to validate that it makes sense for a recommendation engine to group baby clothes with an order of diapers, applesauce and sippy cups.

Câu 6) Give one example of data mining application in the field of healthcare or economic. Based on your example, what kind of data, and data mining method you can use?

Câu 7) Present concept of Clustering and Classification. Take examples to show your ideas.

Clustering

- Clustering is the method of converting a group of abstract objects into classes of similar objects.
- Clustering is a method of partitioning a set of data or objects into a set of significant subclasses called clusters.
- It helps users to understand the structure or natural grouping in a data set and used either as a stand-alone instrument to get a better insight into data distribution or as a pre-processing step for other algorithms

Classification

- Classification is a data mining function that assigns items in a collection to target categories or classes.
- The goal of classification is to accurately predict the target class for each case in the data. Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

Câu 8) Take a pair of attribute that they are highly correlated.

(Strong, Weak) (Long, Short) (Big, Small) (Early, Late)

Câu 9) Do you think "Correlation does imply causality"? and explain why.

A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable. Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events.

Câu 10) What is incomplete (missing) data?

- Data is not always available

  Ex: Many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to:
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry

  - not register history or changes of the data
- Missing data may need to be inferred

Câu 11) How to handle incomplete (missing) data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with:
  - a global constant
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Câu 12) What is noisy data?

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to:
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

Câu 13) How to handle noisy data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries
- Regression
  - smooth by fitting the data into regression functions
- Clustering

    - detect and remove outliers
- Combined computer and human inspection
    - detect suspicious values and check by human (e.g., deal with possible outliers)

CHƯƠNG 2: BÀI TẬP VỀ THUẬT TOÁN APRIORI, ROUGH SET, DECISION TREE, NAIVE BAYESIAN (LAPLACIAN)

Câu 1) Sử dụng Apriori

Coming soon…

Câu 2) Sử dụng Rough Set

B1: Tính $IND_{IS}(B)$. Có nghĩa là dựa vào B={attr1, attr2} liệt kê xem có bao nhiêu loại giá trị khác nhau.

B2: Tính Lower Approximation $\underline{B}X$. Xét từng phần tử trong $IND_{IS}(B)$, các giá trị trong mỗi phần tử đều phải có mặt trong tập X.

B3: Tính Upper Approximation $B\overline{X}$. Xét từng phần tử trong $IND_{IS}(B)$, chỉ cần tối thiểu một giá trị trong mỗi phần tử có mặt trong tập X.

B4: Tính Boundary Region $BR = B\overline{X} – \underline{B}X$

B5: Tính Outsider $B\_OUT = U – B\overline{X}$

B6: Tính Quality coefficient $\alpha_B(X)= \frac{|B\underline{X}|}{|B\overline{X}|}$

- $\alpha_B(X) = 1$ □ X is clear approximation regarding to B
- $\alpha_B(X) < 1$ □ X is rough approximation regarding to B

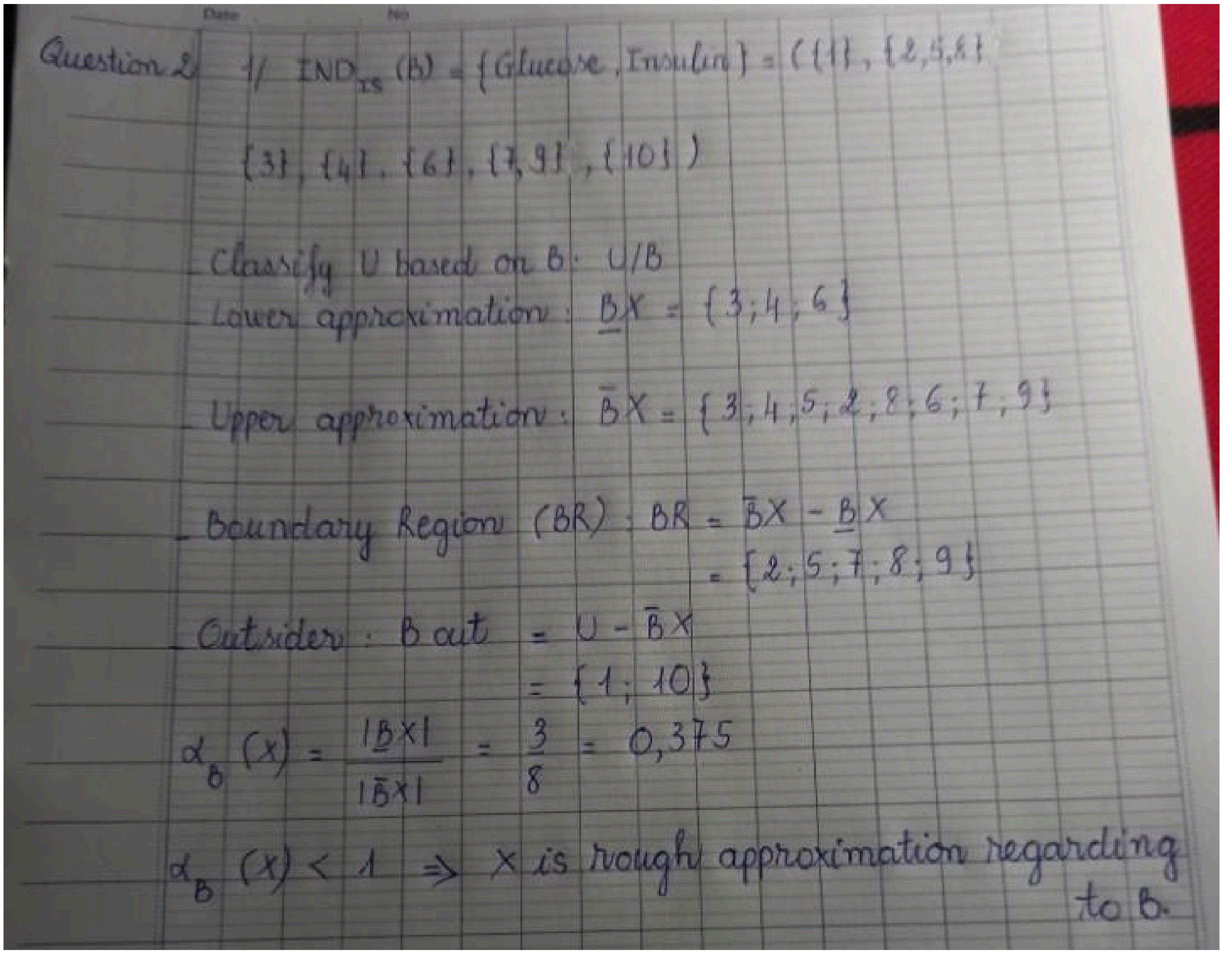
Question 2: (5.5 scores)

Suppose that a diabetes dataset as in the following table (Let *Result* be the decision attribute).

Note: Students can use abbreviations (ex. P for **Pregnancies**) to present the examination.

	Pregnancies (P)	**Glucose (G)**	**Insulin (I)**	**BMI**	**Age (A)**	**Result (R)**
1	0	<70	<15.9	25..30	<18	Negative
2	2	70..92	15.9 ..166	25..30	18..30	Negative
3	3	93..120	>166	>30	31..40	Positive
4	1	>120	>166	>30	>40	Positive
5	2	70..92	15.9 ..166	25..30	18..30	Positive
6	2	93..120	15.9 ..166	25..30	>40	Positive
7	0	>120	15.9 ..166	>30	18..30	Positive
8	1	70..92	15.9 ..166	<25	31..40	Negative
9	2	>120	15.9 ..166	25..30	>40	Negative
10	1	<70	15.9 ..166	>30	>40	Negative

1. Suppose B = {Glucose, Insulin}, X = {3, 4, 5, 6, 7} (***Result*** = "Positive"). Use rough set to compute: upper approximation, lower approximation, and quality coefficient *(1.5 scores)*



Question 2 1/ $IND_{IS}(B) = \{Glucose, Insulin\} = (\{1\}, \{2,5,8\}$

$\{3\}, \{4\}, \{6\}, \{7,9\}, \{10\})$

_ Classify U based on B: U/B

_ Lower approximation: $\underline{B}X = \{3;4;6\}$

_ Upper approximation: $\overline{B}X = \{3;4;5;2;8;6;7;9\}$

_ Boundary Region (BR): $BR = \overline{B}X - \underline{B}X$

$= \{2;5;7;8;9\}$

_ Outsider: B out $= U - \overline{B}X$

$= \{1; 10\}$

$$\alpha_B(x) = \frac{|\underline{B}X|}{|\overline{B}X|} = \frac{3}{8} = 0{,}375$$

$\alpha_B(x) < 1 \Rightarrow$ X is rough approximation regarding to B.

Câu 3) Sử dụng Decision Tree (Gini Index)

B1: Xét từng thuộc tính (ngoại trừ thuộc tính quyết định), xem coi là ở mỗi thuộc tính có bao nhiêu loại giá trị.

B2: Tính Gini của từng loại giá trị. Ta sẽ xét xem loại giá trị đó với thuộc tính quyết định.

$$\text{Gini}(S_{\text{loaiGtri1}}) = 1 - \left(\frac{\textit{Số lượng loại gtrị 1 xét theo thuộc tính quyết định 1}}{\textit{Số lượng loại gtri 1}}\right)^2 - \left(\frac{\textit{Số lượng loại gtri 1 xét theo thuộc tính quyết định 2}}{\textit{Số lượng loại gtri 1}}\right)^2$$

$\text{Gini}(S_{\text{loaiGtri2}}) = \ldots$

$\text{Gini}(S_{\text{loaiGtriN}}) = \ldots$

B3: Sau khi đã tính xong Gini của từng loại giá trị. Ta sẽ tính tiếp tới Gini của thuộc tính đó.

$$\text{Gini}_{\text{ThuocTinh}}(S) = \frac{\textit{Số lượng loại gtri 1}}{\textit{Tổng số dòng}} * \text{Gini}(S_{\text{loaiGtri1}}) + \frac{\textit{Số lượng loại gtri2}}{\textit{Tổng số dòng}} * \text{Gini}(S_{\text{loaiGtri2}}) + \ldots$$

B4: Lặp lại các bước trên cho đến khi ko còn thuộc tính nào để xét.

B5: Chọn ra thuộc tính nào mà có giá trị $\text{Gini}_{\text{ThuocTinh}}(S)$ là NHỎ NHẤT để kết luận.

□ *“ThuocTinh is selected as the root (Gini Index is the minimal value)”*

Question 2: (5.5 scores)

Suppose that a diabetes dataset as in the following table (Let *Result* be the decision attribute).

Note: Students can use abbreviations (ex. P for **Pregnancies**) to present the examination.

	Pregnancies (P)	**Glucose (G)**	**Insulin (I)**	**BMI**	**Age (A)**	**Result (R)**
1	0	<70	<15.9	25..30	<18	Negative
2	2	70..92	15.9 ..166	25..30	18..30	Negative
3	3	93..120	>166	>30	31..40	Positive
4	1	>120	>166	>30	>40	Positive
5	2	70..92	15.9 ..166	25..30	18..30	Positive
6	2	93..120	15.9 ..166	25..30	>40	Positive
7	0	>120	15.9 ..166	>30	18..30	Positive
8	1	70..92	15.9 ..166	<25	31..40	Negative
9	2	>120	15.9 ..166	25..30	>40	Negative
10	1	<70	15.9 ..166	>30	>40	Negative

1. Suppose B = {Glucose, Insulin}, X = {3, 4, 5, 6, 7} (***Result*** = “Positive”). Use rough set to compute: upper approximation, lower approximation, and quality coefficient *(1.5 scores)*
2. Determine the root of Decision Tree using Gini Index. *(2.0 scores)*

2/ Gini Index

* Pregnancies (P)

$$\text{Gini}(S_0) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0,5$$

$$\text{Gini}(S_1) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \approx 0,44$$

$$\text{Gini}(S_2) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 \approx 0,5$$

$$\text{Gini}(S_3) = 1 - \left(\frac{1}{1}\right)^2 = 0$$

$$\Rightarrow \text{Gini}_{\text{Pregnancies}}(S) = \frac{2}{10} \times 0,5 + \frac{3}{10} \times 0,44 + \frac{4}{10} \times 0,5 + \frac{1}{10} \times 0 = 0,432$$

* Glucose (G)

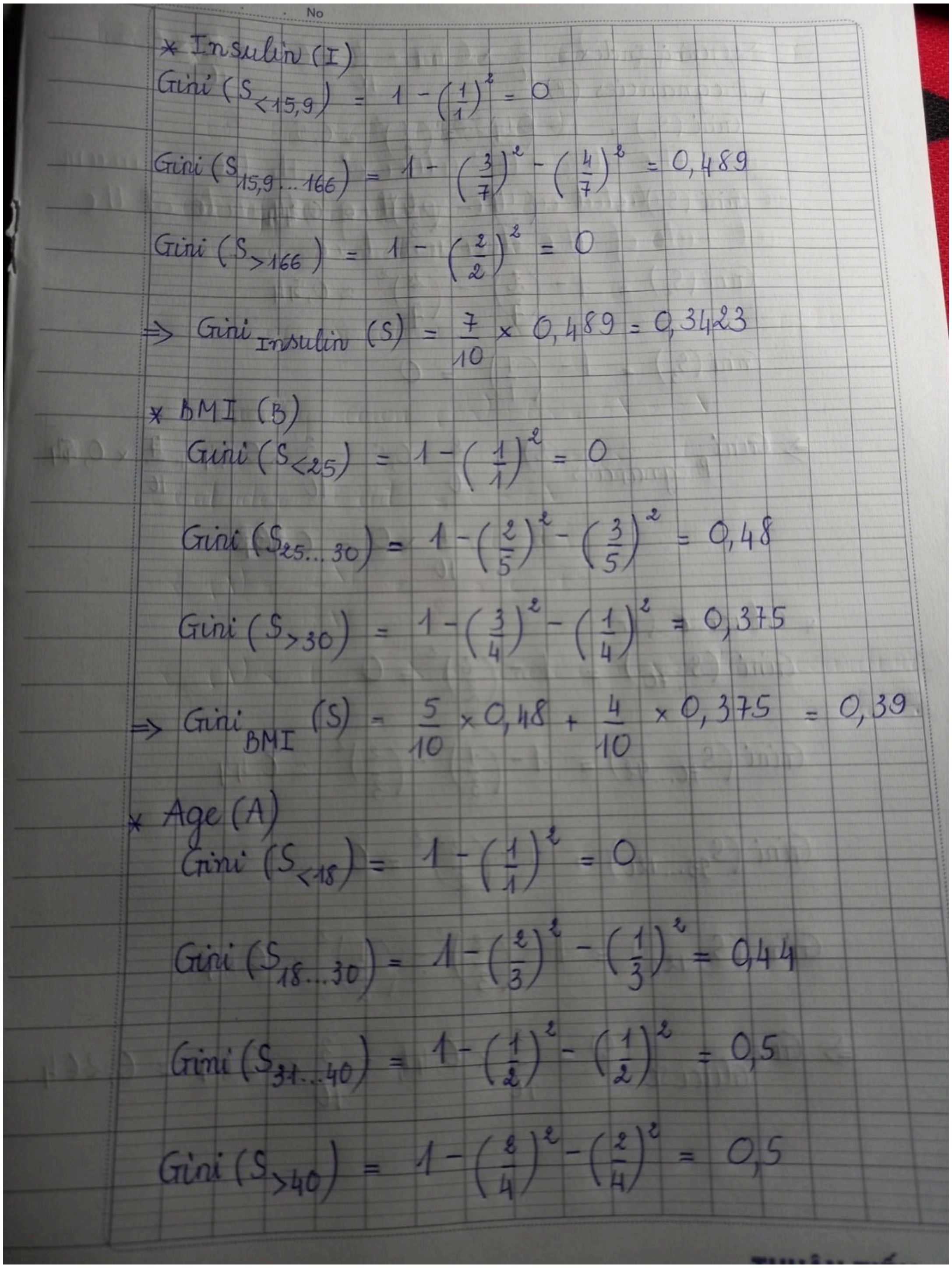
$$\text{Gini}(S_{<70}) = 1 - \left(\frac{2}{2}\right)^2 = 0$$

$$\text{Gini}(S_{70\dots92}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \approx 0,44$$

$$\text{Gini}(S_{93\dots120}) = 1 - \left(\frac{2}{2}\right)^2 = 0$$

$$\text{Gini}(S_{>120}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \approx 0,44$$

$$\Rightarrow \text{Gini}_{\text{Glucose}}(S) = \frac{3}{10} \times 0,44 + \frac{3}{10} \times 0,44 = 0,264$$



No

* Insulin (I)

$$\text{Gini}(S_{<15,9}) = 1 - \left(\frac{1}{1}\right)^2 = 0$$

$$\text{Gini}(S_{15,9\ldots166}) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0,489$$

$$\text{Gini}(S_{>166}) = 1 - \left(\frac{2}{2}\right)^2 = 0$$

$$\Rightarrow \text{Gini}_{\text{Insulin}}(S) = \frac{7}{10} \times 0,489 = 0,3423$$

* BMI (B)

$$\text{Gini}(S_{<25}) = 1 - \left(\frac{1}{1}\right)^2 = 0$$

$$\text{Gini}(S_{25\ldots30}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0,48$$

$$\text{Gini}(S_{>30}) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0,375$$

$$\Rightarrow \text{Gini}_{\text{BMI}}(S) = \frac{5}{10} \times 0,48 + \frac{4}{10} \times 0,375 = 0,39$$

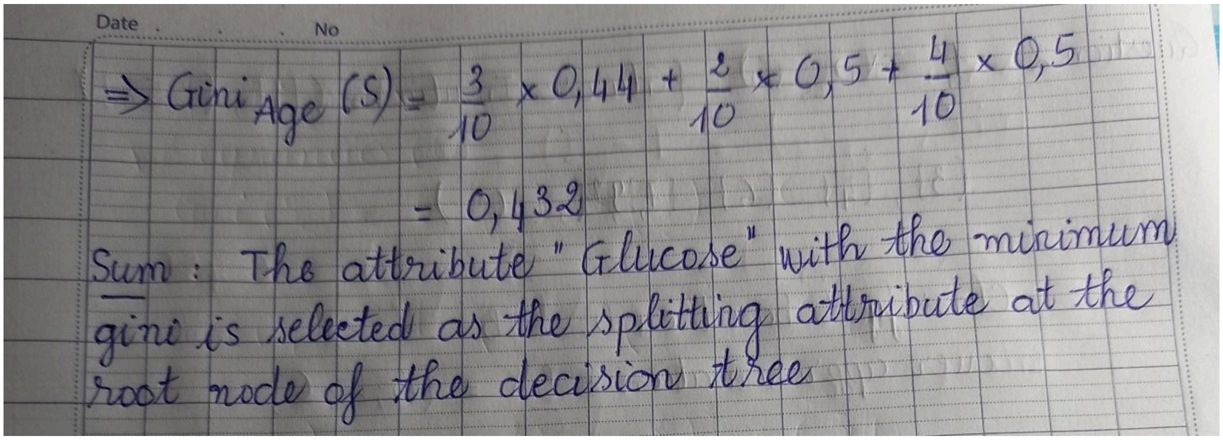
* Age (A)

$$\text{Gini}(S_{<18}) = 1 - \left(\frac{1}{1}\right)^2 = 0$$

$$\text{Gini}(S_{18\ldots30}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0,44$$

$$\text{Gini}(S_{31\ldots40}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0,5$$

$$\text{Gini}(S_{>40}) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0,5$$



Date . . . No

$\Rightarrow Gini_{Age}(S) = \frac{3}{10} \times 0,44 + \frac{2}{10} \times 0,5 + \frac{4}{10} \times 0,5$

$= 0,432$

Sum : The attribute "Glucose" with the minimum gini is selected as the splitting attribute at the root node of the decision tree

Câu 4) Sử dụng Naive Bayesian (Laplacian)

B1: Ta chia làm hai trường hợp:

- TH1 là dựa trên GIÁ TRỊ THỨ NHẤT của thuộc tính quyết định.

  vd: P(Result=Positive | X) = $\frac{P(X\ |Result=Positive)*\ P(Result=Positive)}{P(X)}$

- TH2 là dựa trên GIÁ TRỊ CÒN LẠI của thuộc tính quyết định.

  vd: P(Result=Negative | X) = $\frac{P(X\ |Result=Negative)*\ P(Result=Negative)}{P(X)}$

B2: Tính xác suất các giá trị của thuộc tính quyết định:

$$P(gtriTTQD) = \frac{soLanXuatHienCuaTTQD+\ 1}{soDong+\ soLuongLoaiGtri}$$

B3: Tính xác suất của X dựa trên các giá trị của thuộc tính quyết định:

$$P(X\ |\ gtriTTQD) = \frac{soLanXuatHienCuaXDoiVoiTTQD+1}{soDongTTQD+\ soLuongLoaiGtriX}$$

B4: Tính P(X | gtriTTQD) * P (gtriTTQD) bằng cách ta nhân gtri thu đc ở B2 với các gtri thu đc ở B3

B5: Kết luận. Ta thấy nếu P(Result=Positive | X) > P (Result=Negative | X) “There for X belongs to class Result=Positive”. Và ngược lại.

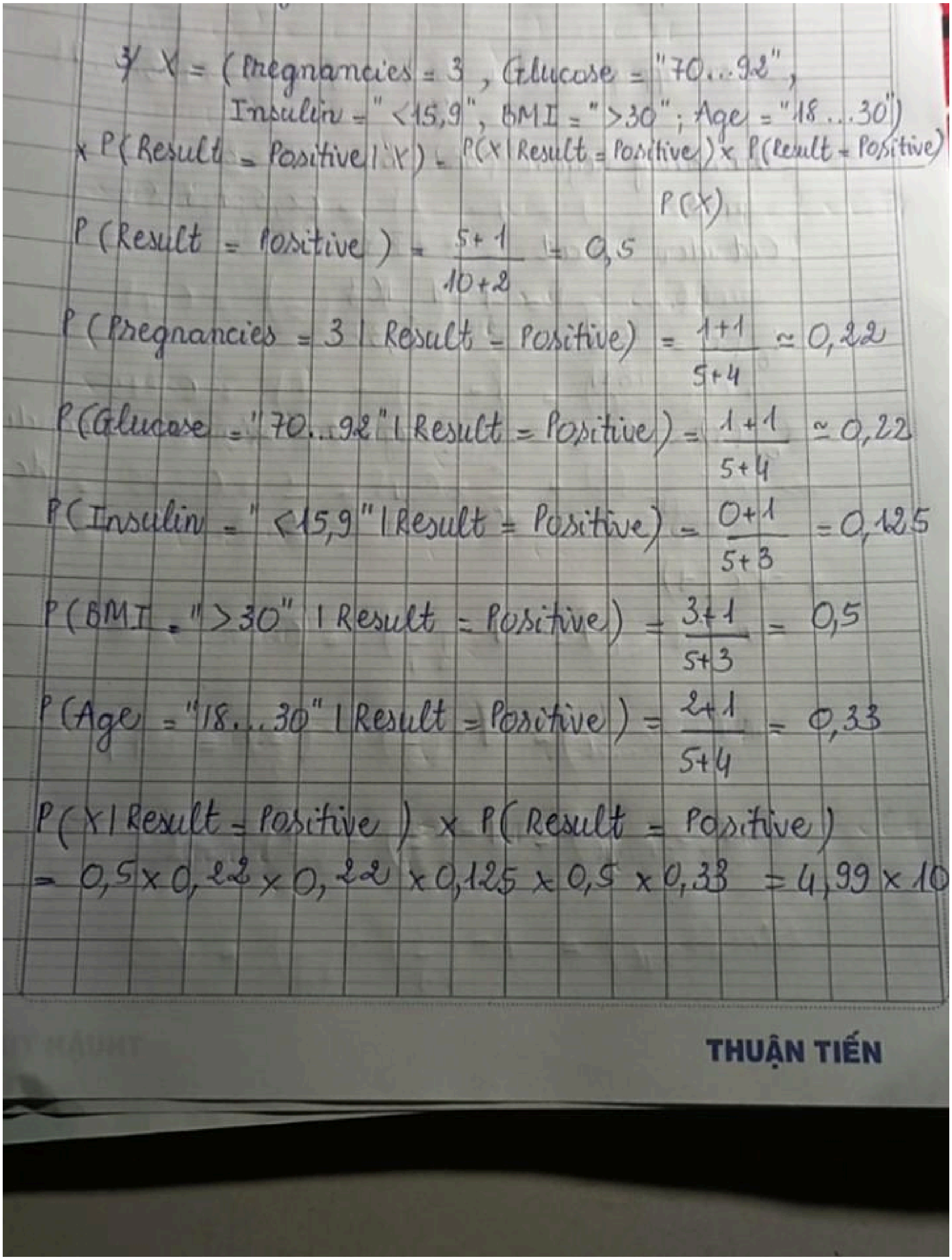
Question 2: (5.5 scores)

Suppose that a diabetes dataset as in the following table (Let *Result* be the decision attribute).

Note: Students can use abbreviations (ex. P for **Pregnancies**) to present the examination.

	Pregnancies (P)	**Glucose (G)**	**Insulin (I)**	**BMI**	**Age (A)**	**Result (R)**
1	0	<70	<15.9	25..30	<18	Negative
2	2	70..92	15.9 ..166	25..30	18..30	Negative
3	3	93..120	>166	>30	31..40	Positive
4	1	>120	>166	>30	>40	Positive
5	2	70..92	15.9 ..166	25..30	18..30	Positive
6	2	93..120	15.9 ..166	25..30	>40	Positive
7	0	>120	15.9 ..166	>30	18..30	Positive
8	1	70..92	15.9 ..166	<25	31..40	Negative
9	2	>120	15.9 ..166	25..30	>40	Negative
10	1	<70	15.9 ..166	>30	>40	Negative

1. Suppose B = {Glucose, Insulin}, X = {3, 4, 5, 6, 7} (***Result*** = "Positive"). Use rough set to compute: upper approximation, lower approximation, and quality coefficient *(1.5 scores)*
2. Determine the root of Decision Tree using Gini Index. *(2.0 scores)*
3. Given a sample *X = (Pregnancies* **=** *3, Glucose = "70..92", Insulin = "<15.9", BMI = ">30"; Age = 18..30")*, what would a Naïve Bayesian classification using Laplacican correction of the ***Result*** for sample X be? *(2.0 scores)*



3/ X = (Pregnancies = 3, Glucose = "70...92",
Insulin = "<15,9", BMI = ">30"; Age = "18...30")

$$\times P(\text{Result} = \text{Positive} \mid X) = \frac{P(X \mid \text{Result} = \text{Positive}) \times P(\text{Result} = \text{Positive})}{P(X)}$$

$$P(\text{Result} = \text{Positive}) = \frac{5+1}{10+2} = 0{,}5$$

$$P(\text{Pregnancies} = 3 \mid \text{Result} = \text{Positive}) = \frac{1+1}{5+4} \simeq 0{,}22$$

$$P(\text{Glucose} = \text{"70...92"} \mid \text{Result} = \text{Positive}) = \frac{1+1}{5+4} \simeq 0{,}22$$

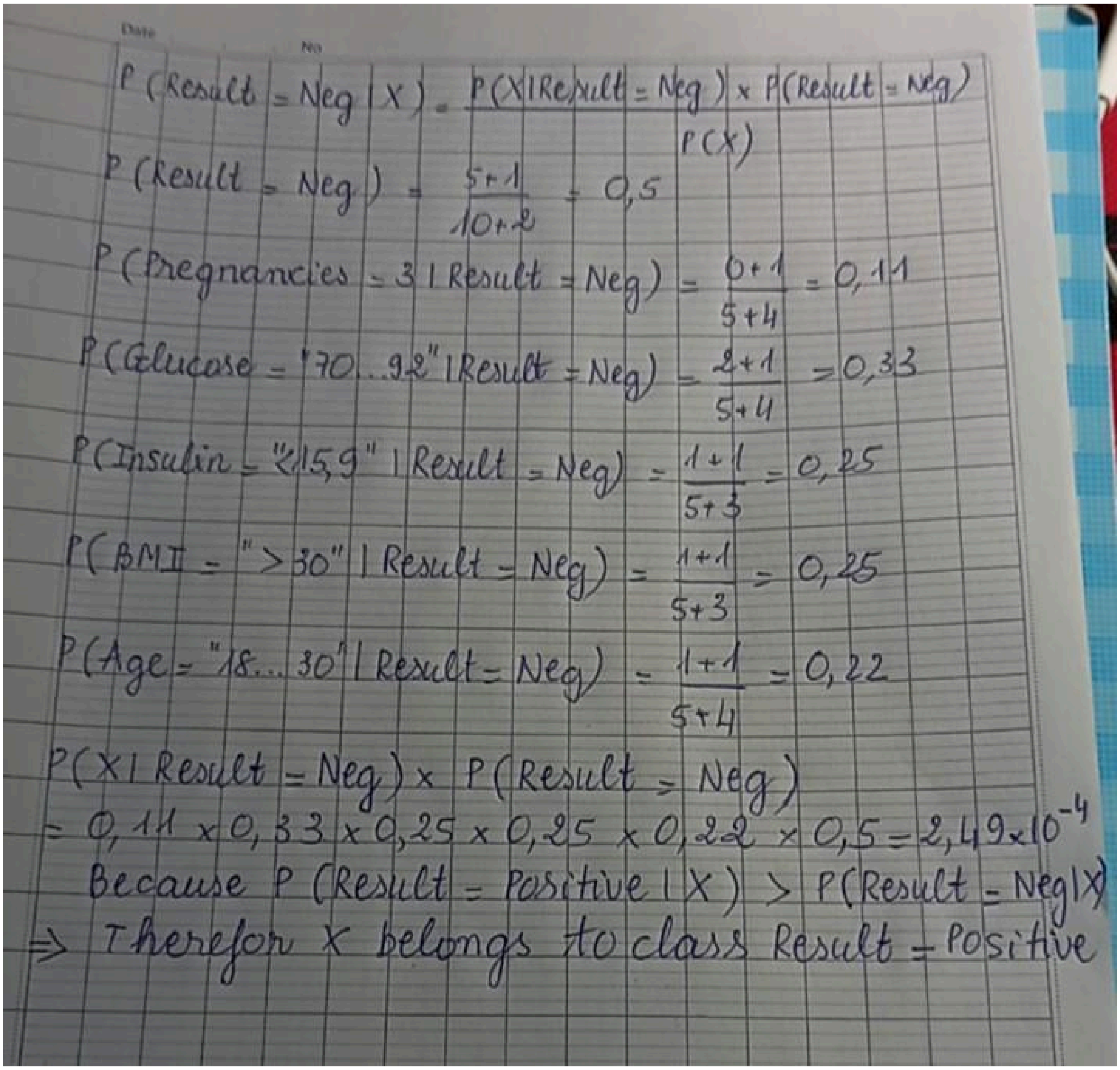
$$P(\text{Insulin} = \text{"<15,9"} \mid \text{Result} = \text{Positive}) = \frac{0+1}{5+3} = 0{,}125$$

$$P(\text{BMI} = \text{">30"} \mid \text{Result} = \text{Positive}) = \frac{3+1}{5+3} = 0{,}5$$

$$P(\text{Age} = \text{"18...30"} \mid \text{Result} = \text{Positive}) = \frac{2+1}{5+4} = 0{,}33$$

$P(X \mid \text{Result} = \text{Positive}) \times P(\text{Result} = \text{Positive})$

$= 0{,}5 \times 0{,}22 \times 0{,}22 \times 0{,}125 \times 0{,}5 \times 0{,}33 = 4{,}99 \times 10$



Date

No

$$P(\text{Result} = \text{Neg} \mid X) = \frac{P(X \mid \text{Result} = \text{Neg}) \times P(\text{Result} = \text{Neg})}{P(X)}$$

$$P(\text{Result} = \text{Neg}) = \frac{5+1}{10+2} = 0{,}5$$

$$P(\text{Pregnancies} = 3 \mid \text{Result} = \text{Neg}) = \frac{0+1}{5+4} = 0{,}11$$

$$P(\text{Glucose} = \text{"70...92"} \mid \text{Result} = \text{Neg}) = \frac{2+1}{5+4} = 0{,}33$$

$$P(\text{Insulin} = \text{"<15,9"} \mid \text{Result} = \text{Neg}) = \frac{1+1}{5+3} = 0{,}25$$

$$P(\text{BMI} = \text{">30"} \mid \text{Result} = \text{Neg}) = \frac{1+1}{5+3} = 0{,}25$$

$$P(\text{Age} = \text{"18...30"} \mid \text{Result} = \text{Neg}) = \frac{1+1}{5+4} = 0{,}22$$

$$P(X \mid \text{Result} = \text{Neg}) \times P(\text{Result} = \text{Neg})$$

$$= 0{,}11 \times 0{,}33 \times 0{,}25 \times 0{,}25 \times 0{,}22 \times 0{,}5 = 2{,}49 \times 10^{-4}$$

Because $P(\text{Result} = \text{Positive} \mid X) > P(\text{Result} = \text{Neg} \mid X)$

⇒ Therefor X belongs to class Result = Positive

CHƯƠNG 3: BÀI TẬP VỀ THUẬT TOÁN K-MEANS

B1: Tính Centroid Vector (Vecto trọng tâm). Ta nhìn vào cái bảng mà đề cho, ta xét từng dòng Ci, nhìn vào những ô có đánh số “1”. Ta sẽ tính trung bình cộng những tọa độ đc đánh số “1”.

B2: Tính khoảng cách từ các tọa độ đến từng cụm Ci. Sử dụng công thức Euclid:

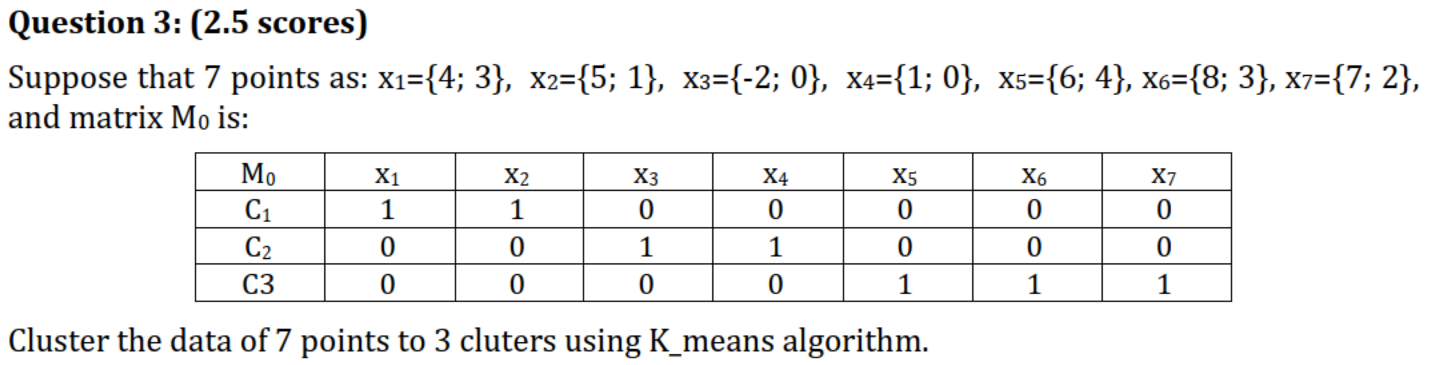
$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

B3: Ta kẻ một cái bảng, cột đầu tiên lần lượt là các cụm Ci, dòng đầu tiên sẽ là tọa độ của các điểm.

B4: Ta sẽ so sánh khoảng cách của từng tọa độ đến các cụm Ci. Gtri nào nhỏ nhất thì cái tọa độ mà ta đang xét sẽ thuộc về cụm tương ứng. Và ta sẽ đánh số “1” tại tọa độ đó, những cụm còn lại sẽ đánh số 0.

B5: Dấu hiệu kết thúc

- Nếu như cái bảng mà ta vẽ ở B4 mà trùng với cái bảng mà đề cho □ Dừng □ Kết thúc bài.
- Nếu như đề yêu cầu “Show matrix U1” thì ta vẽ xong cái bảng ở B4 □ Dừng □ Kết thúc bài.
- Nếu đề ko yêu cầu gì hết thì sau khi hoàn thành B4 mà cái bảng ta làm ko trùng vs đề bài □ Quay lại B1, làm lại các bước B1 B2 B3 B4 dựa trên cái bảng ta vừa vẽ □ Lặp lại cho đến khi nào bảng ta vẽ trùng với đề bài thì thôi.

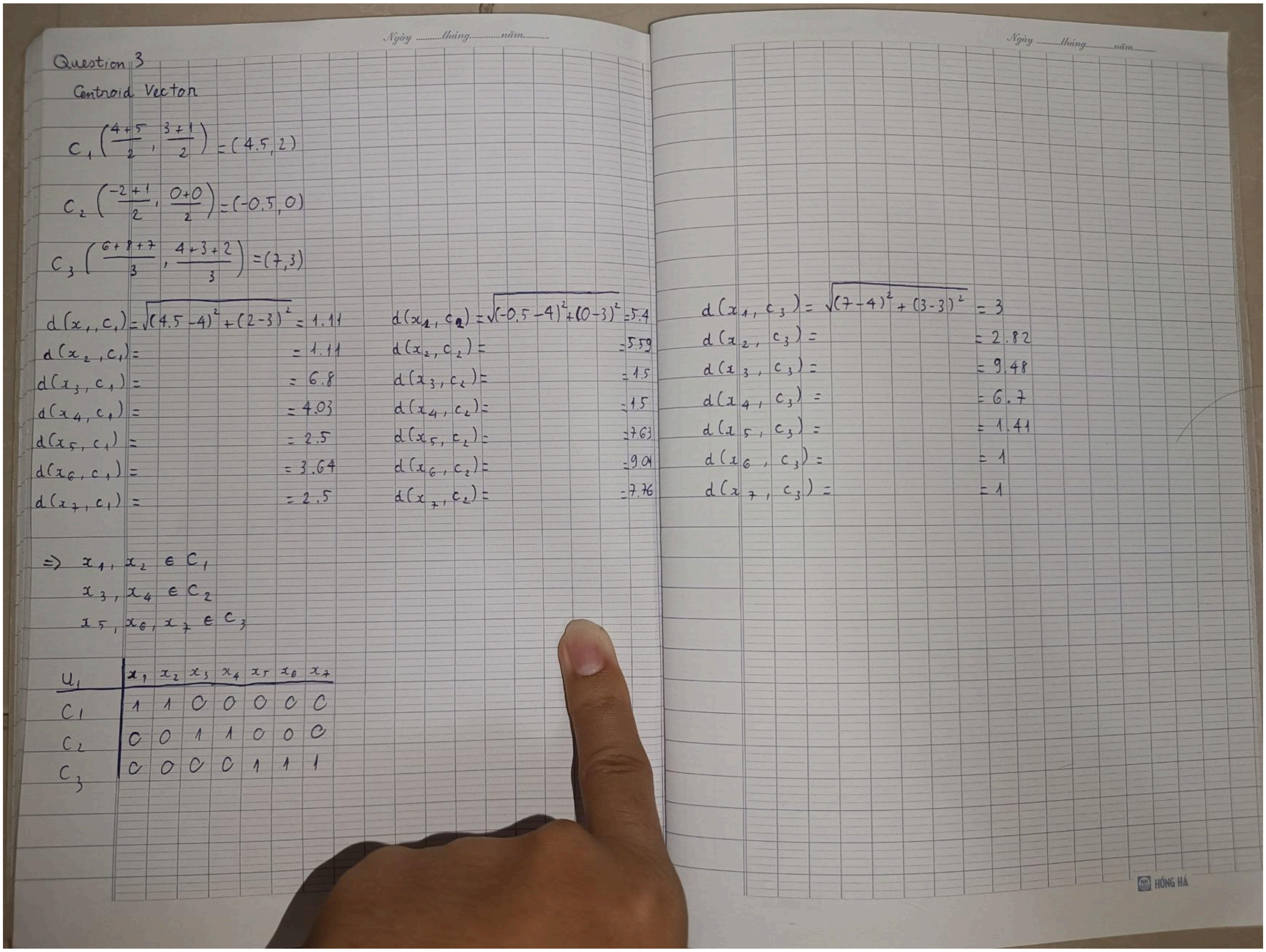


Question 3: (2.5 scores)

Suppose that 7 points as: x_1={4; 3}, x_2={5; 1}, x_3={-2; 0}, x_4={1; 0}, x_5={6; 4}, x_6={8; 3}, x_7={7; 2}, and matrix M_0 is:

M_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
C_1	1	1	0	0	0	0	0
C_2	0	0	1	1	0	0	0
C3	0	0	0	0	1	1	1

Cluster the data of 7 points to 3 cluters using K_means algorithm.



Question 3

Centroid Vector

$$c_1\left(\frac{4+5}{2}, \frac{3+1}{2}\right) = (4.5, 2)$$

$$c_2\left(\frac{-2+1}{2}, \frac{0+0}{2}\right) = (-0.5, 0)$$

$$c_3\left(\frac{6+8+7}{3}, \frac{4+3+2}{3}\right) = (7, 3)$$

$d(x_1, c_1) = \sqrt{(4.5-4)^2 + (2-3)^2} = 1.11$

$d(x_2, c_1) = \quad = 1.11$

$d(x_3, c_1) = \quad = 6.8$

$d(x_4, c_1) = \quad = 4.03$

$d(x_5, c_1) = \quad = 2.5$

$d(x_6, c_1) = \quad = 3.64$

$d(x_7, c_1) = \quad = 2.5$

$d(x_1, c_2) = \sqrt{(-0.5-4)^2 + (0-3)^2} = 5.4$

$d(x_2, c_2) = \quad = 5.59$

$d(x_3, c_2) = \quad = 1.5$

$d(x_4, c_2) = \quad = 1.5$

$d(x_5, c_2) = \quad = 7.63$

$d(x_6, c_2) = \quad = 9.01$

$d(x_7, c_2) = \quad = 7.76$

$d(x_1, c_3) = \sqrt{(7-4)^2 + (3-3)^2} = 3$

$d(x_2, c_3) = \quad = 2.82$

$d(x_3, c_3) = \quad = 9.48$

$d(x_4, c_3) = \quad = 6.7$

$d(x_5, c_3) = \quad = 1.41$

$d(x_6, c_3) = \quad = 1$

$d(x_7, c_3) = \quad = 1$

$\Rightarrow x_1, x_2 \in C_1$

$x_3, x_4 \in C_2$

$x_5, x_6, x_7 \in C_3$

U_1	x_1	x_2	x_3	x_4	x_5	x_6	x_7
C_1	1	1	0	0	0	0	0
C_2	0	0	1	1	0	0	0
C_3	0	0	0	0	1	1	1