AI & Technology Department – Data Scientist Position

Dear Candidate,

Thanks for your interest in the Data Scientist job at ' ompany. For the position of Data Scientist, we have some tests for you. The test is divided into two parts:

- Probability & Statistics
- Data & Coding

Good luck!

Probability & Statistics questions:

1. The probability of passing three candidates A, B, and C in one interview are 60%, 10%, and 20% respectively. Find the probability of success of at least two *(2 points)*
   p = 0,6*0,1*0,8 + 0,6*0,2*0,9 + 0,1*0,2*0,4 + 0,6*0,2*0,1

2. Nine cards are drawn from a deck of 52 cards. What is the probability that 5 cards are diamonds and 2 are hearts and 2 are kings? *(2 points)*
   king diamond and king heart: 4C12*12*2C45

3. There are transactions from banks A, B, and C at your company. About 30% of the population comes from A, 20% from B, and 50% from C. Knowing that there is a chance of fraud in the transaction history. The percentages for banks A, B, and C are 3%, 5%, and 4%, respectively.

Which bank will most likely belong in the event of fraud detection? *(2 points)*

4. Assume that the amount of transaction in Service X in your company follow the Normal distribution with mean μ. A random sample of 250 000 transactions yielded an average amount of $\bar{x}$ = 525 000 VND, and the standard deviation of the 250 000 transactions is found to be s = 28 340. A 90% confidence interval for μ is *(Hint: The t critical value for 90% CI for Degrees of freedom = 24 is 1.711*) *(2 points)*

Data & Coding:

Please download the Kaggle dataset:

https://www.kaggle.com/datasets/vipin20/transaction-data and follow the instruction below

1. Clean data and filter Outliers (*Explain your ideas and ways how to find out Outliers and the ways to treat them*) (*2 points*)
2. Calculate the number of Items purchased and prices in each month (*1 point*)
3. Report about the efficiency of selling in each time frame (*1 point*)
4. Calculate the number of items purchased for each userID in 30 days for each day (*Add new column*) (*2 points*)

For example:

UserID	…	TransactionTime	NumberOfItemsPurchased	NumberOfItemsPurchased_30days
ABC		1, March 2022	6	6
ABC		15, March 2022	12	18
ABC		28, March 2022	20	38
ABC		14, April 2022	10	42
ABC		30, April 2022	50	60

5. With the goal of purchasing more items, write ideas using Machine Learning algorithms for a recommendation system in this data. Implement your ideas (*Explain the reason why you propose these ideas and prove their efficiency*) (*6 points*)

This is the end of the test! Thank you so much for your efforts!