**2021 May Day MCM**

Problem C. Data-driven anomaly detection and warning problem

To promote the high-quality development of production enterprises, the most fundamental bottom line is to ensure safety and prevent risks. The data generated in the production process can reflect the potential risks in real time. Appendix 1 includes the time series data recorded by the equipments in the production area from 00:00:00 to 22:59:59 on a certain day in a production enterprise (data desensitization has been carried out). The specific name of the data is not given in this question. These data may be temperature, concentration, pressure and other data closely related to safety issue.

Please use the data in Appendix 1 and reply the following questions.

**Question 1:** The data presented in Appendix 1 may fluctuate. All the fluctuations are within the safety threshold. Some fluctuations may be normal fluctuations, such as those changes caused by external temperature or production, or sensor false alarms. Such fluctuations are regular, independent and sporadic, and will not produce safety risks. We regard them as non-risky anomalies and do not need to intervene with them. Some fluctuations are characterized by persistence and linkage, etc. The occurrence of these fluctuations is caused by unstable factors in the production process, which indicates that there may be hidden unsafe hazards. Therefore, we regard these fluctuations as risky anomalies, which require human intervention, analysis and assessment of risk levels. Please establish a mathematical model to determine non-risky anomaly data and risky anomaly data.

**Question 2:** Using the results of question 1, please establish a risky anomaly detection model, and give the quantitative evaluation method of the abnormal degree of risky anomaly data. Please use score (0-100 points) to evaluate the abnormal degree of the whole data (all sensors) at each moment (the higher the score is, the higher the abnormal degree is). You need to apply the established model and the data in Appendix 1 to find the top 5 moments with the highest abnormal score in the data and the corresponding abnormal sensor numbers of these 5 moments (Only 5 sensor numbers with the highest abnormal degree are needed to be filled in at each moment, and if the number of abnormal sensors is less than 5, fill them in the table as much as possible; If the score is 0, you don't need to fill in the abnormal sensor number), and give relevant methods to evaluate the results you give.

Table 1 Results for Question 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | The first  high score | The second  high score | The third  high score | The forth  high score | The fifth  high score |
| Score of abnormal degree |  |  |  |  |  |
| Number of abnormal time |  |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |  |

**Question 3:** In order to find the potential risks in the future production process in advance, please establish a risky anomaly warning model to predict the risky anomaly that may occur from 23:00:00 to 23:59:59 on the day. Using the quantitative evaluation method of risky anomaly degree given in Question 2, please find the highest abnormal score and corresponding abnormal sensor number of each time period in 23:00:00 to 23:59:59 (see Table 2). (Only 5 sensor numbers with the highest abnormal degree are needed to be filled in at each moment, and if the number of abnormal sensors is less than 5, fill them in the table as much as possible; If the score is 0, you don't need to fill in the abnormal sensor number).

Table 2 Results for Question 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time | 23:00:00-23:14:59 | 23:15:00-23:29:59 | 23:30:00-23:44:59 | 23:45:00-23:59:59 |
| Highest score of abnormal data |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |
| Number of abnormal sensor |  |  |  |  |

**Question 4:** According to the results in questions 2 and 3, please establish a mathematical model to evaluate the safety of the entire production system of the production enterprise. Please score the safety from 0 to 100 every 30 minutes from 00:00:00 to 23:59:59, with 0 indicating the lowest safety and 100 indicating the highest safety (including the score from 00:00:00 to 23:00:00 and the predicted score from 23:00:00 to 23:59:59). And please use appropriate method to evaluate the results and conduct sensitivity analysis.

Table 3 Results for Question 4

|  |  |  |  |
| --- | --- | --- | --- |
| Time | Safety score | Time | Safety score |
| 00:30:00 |  | 12:30:00 |  |
| 01:00:00 |  | 13:00:00 |  |
| 01:30:00 |  | 13:30:00 |  |
| 02:00:00 |  | 14:00:00 |  |
| 02:30:00 |  | 14:30:00 |  |
| 03:00:00 |  | 15:00:00 |  |
| 03:30:00 |  | 15:30:00 |  |
| 04:00:00 |  | 16:00:00 |  |
| 04:30:00 |  | 16:30:00 |  |
| 05:00:00 |  | 17:00:00 |  |
| 05:30:00 |  | 17:30:00 |  |
| 06:00:00 |  | 18:00:00 |  |
| 06:30:00 |  | 18:30:00 |  |
| 07:00:00 |  | 19:00:00 |  |
| 07:30:00 |  | 19:30:00 |  |
| 08:00:00 |  | 20:00:00 |  |
| 08:30:00 |  | 20:30:00 |  |
| 09:00:00 |  | 21:00:00 |  |
| 09:30:00 |  | 21:30:00 |  |
| 10:00:00 |  | 22:00:00 |  |
| 10:30:00 |  | 22:30:00 |  |
| 11:00:00 |  | 23:00:00 |  |
| 11:30:00 |  | 23:30:00 |  |
| 12:00:00 |  | 23:59:59 |  |