DATA ANALYSIS COMPETITION 2023

QUESTION SHEET
DAC 2023

PRELIMINARY ROUND

GUIDELINE PRELIMINARY ROUND

1. Participants, who are registered as teams consist of 2 people from the same university and as the active students, will compete in the preliminary round.
2. Participants pay for the account activation fee to activate their registered accounts.
3. Participants with the activated accounts may access the question via PRS website during the preliminary round. In addition, the leader of each team will also receive questions and datasets via email as a backup.
4. Participants are allowed using any programming language.
5. The time period to accomplish and collect the answers is 6 days, from September 26th 2023 until October 1st 2023.
6. Participants are not allowed to submit the answers exceeds the allocated time. The subject of answer sheet is : Registration Number_Team's name_University.
7. Participants who qualify for the semifinal round are 15 teams. The semifinalists are selected with rank based selection, details of rank based selection.

GENERAL INSTRUCTIONS PRELIMINARY ROUND

1. Write your results of the analysis on the answer sheet we provide. In addition, we require you to:
   a. Write your analysis completely and clearly.
   b. Write a report with a maximum of 6 pages in English. Use paper size A4, font Times New Roman size 14 for chapter titles and the rest use size 12, with line spacing 1.15, and margin 3.5 3 3 3, following the provided template.
   c. Participants should include the syntax, additional figures, and references in the attachment page right after the article. The attachment page are not included in the previous 6 pages and has no page limit.
   d. Images that are not of paramount importance can be included on the attachment page to conserve page usage. The attachment page has no page limit. Ensure that you index the images you attach to the attachment page.
2. Save the answer sheet with the format file:
   a. For the article is pdf and docx.
   b. For the predictions is csv, with the format name: Regristration Number_ Team's name_University.
3. All answers must be compressed in a RAR format with the format name: Registration Number_Team's name_University.rar.
   Example: DAC-01-2581_PRS2023_ITS.rar (All participants may check the registration number at the pre-registration webpage on PRS Website).
4. Upload your team's answer maximum on October 1st 2023 at 23.59 Western Indonesian Time (GMT+7) on the PRS website. If participants send in format which does not match, then the participant's answer sheet will not be corrected.
5. Participants are allowed using any programming language but not allowed using excel only.
6. All kinds of cheating, including copy answers from other teams will result in disqualification.
7. The judge's decision cannot be contested.
8. Any participant who breaks the general instructions will be disqualified.
9. The proportion of assessments in the preliminary round is as follow:
   a. CHAPTER I: Introduction (10%)
   b. CHAPTER II: Theoritical Frame Work (20%)
   c. CHAPTER III: Analytical Steps (10%)
   d. CHAPTER IV: Analysis of Result (35%)
   e. CHAPTER V: Conclusion and Recommendation (15%)
   f. Writing (10%)
10. If there are any problems or questions during the test, please contact:
    a. Galih (Whatsapp: +62 81257623853/Line: galihfx)
    b. Naufal (Whatsapp: +62 88217786396/Line: nlt160303)

Data Analysis Competition (DAC) 2023

At this moment, Information technology is developing rapidly and causing an increase in the delivery of data and global information. The telecommunications sector has also snowballed and undergone significant transformations in the last few decades, especially in Indonesia. Telecommunications has become an important sector in Indonesia's development. The development of telecommunications has significantly impacted increasing connectivity, access to information, and community empowerment. The Indonesian government has made various efforts to advance the telecommunications sector in this country in an attempt to realize the vision of "Indonesia Emas 2045". Increasing speed and capacity in terms of internet connectivity and data transfer is a form of telecommunications development and the development of wireless networks that change how we communicate and connect with the world. Telecommunications is a field that includes sending information over long distances using technology. Telecommunications development continues, focusing on speed and better connectivity in many networks.

However, along with the development of telecommunications, a network security challenge emerged that was quite threatening, namely network attacks. Network attacks refer to unauthorized or malicious attempts to manipulate, damage, or access computer systems, networks, or communications infrastructure without permission. These network attacks significantly impact a variety of other threats that can result in the leakage of sensitive and confidential information. Network attacks can risk destroying the owner's information assets without adequate protection. Therefore, to prevent it, it is necessary to detect attacks on the network with software development so that proper handling can occur. In this case, we already have data about network traffic and types of attacks on the web that log before.

No.	Variable Name	Description
1	Duration	The length of time (in seconds) of the connection.
2	Protocol_type	The protocol type of the connection.
3	Service	Name of the network service on the target host, such as HTTP, FTP, etc.
4	Flag	A flag field that represents the status of the connection (e.g., SF for a normal connection, REJ for rejected connection attempts).
5	Src_bytes	The number of bytes transferred from the source to the destination.

6	Dst_bytes	The number of bytes transferred from the destination to the source.
7	Land	Indicates whether the connection is from/to the same host/port (1 if yes, 0 if no).
8	Wrong_fragment	The number of "wrong" fragments in a connection.
9	Urgent	The number of urgent packets in a connection.
10	Hot	The number of "hot" indicators in a connection.
11	Num_failed_logins	The number of failed login attempts.
12	Logged_in	Indicates whether a login was successfully performed (1 if yes, 0 if no).
13	Num_compromised	The number of hosts compromised through the connection.
14	Root_shell	Indicates whether a root shell was obtained (1 if yes, 0 if no).
15	Su_attempted	Indicates whether a "su root" command was attempted (1 if yes, 0 if no).
16	Num_root	The number of root access obtained.
17	Num_file_creations	The number of file creation operations.
18	Num_shells	The number of shell prompts.
19	Num_access_files	The number of operations on access control files.
20	Num_outbound_cmds	The number of outbound commands in an FTP session.
21	Is_hot_login	Indicates whether the login belongs to the "hot" list (1 if yes, 0 if no).
22	Is_guest_login	Indicates whether the login is a guest login (1 if yes, 0 if no).

23	Count	The number of connections to the same host as the current connection in the past two seconds.
24	Srv_count	The number of connections to the same service as the current connection in the past two seconds.
25	Serror_rate	The percentage of connections that had "SYN" errors. **(SYN : SYNCHRONIZE)**
26	Srv_serror_rate	The percentage of connections to the same service that had "SYN" errors.
27	Rerror_rate	The percentage of connections that had "REJ" errors. **(REJ : REJECTION)**
28	Srv_rerror_rate	The percentage of connections to the same service that had "REJ" errors.
29	Same_srv_rate	The percentage of connections to the same service.
30	Diff_srv_rate	The percentage of connections to different services.
31	Srv_diff_host_rate	The percentage of connections to different hosts.
32	Dst_host_count	The number of connections to the same destination host in the past two seconds.
33	Dst_host_srv_count	The number of connections to the same destination service in the past two seconds.
34	dst_host_same_srv_rate	The percentage of connections to the same destination service.
35	dst_host_diff_srv_rate	The percentage of connections to different destination services.
36	dst_host_same_src_port_rate	The percentage of connections to the same destination host that originate from the same source port.
37	dst_host_srv_diff_host_rate	The percentage of connections to the same destination service but originating from different host.

38	dst_host_serror_rate	The percentage of connections to a specific destination host that have a “SYN” error.
39	dst_host_srv_serror_rate	The percentage of connections to a specific destination service that have a “SYN” error.
40	dst_host_rerror_rate	The percentage of connections to a specific destination host that have a “REJ” error.
41	dst_host_srv_rerror_rate	The percentage of connections to a specific destination service that have a “REJ” error.
42	type_of_attack	Type of attack.

From the data collected on network traffic, you are a consultant for the development team of network attack detection software asked to develop software to detect types of network attacks based on the characteristics of the network traffic that occurs. If you find inconsistent information about the data related to the description, you must make your own decision and provide reasons why you are taking such action. To report results, you must write in article format (see General Instructions) a maximum of 6 pages and an attachment of your code.

The Questions

1. What are the problems and objectives of the appropriate analysis based on the question narrative?
2. Explain the most suitable method to solve the identified problem!
3. If any, explain the underlying assumptions of chosen method!
4. Explain the step-by-step process involved to analyze the data!
5. Provide justifications for any necessary actions in preparing data for analysis!
6. Explore the details of imporatant information that would help analyze the dataset further!
7. Show and explain your result and analysis!
8. Which variables within the data set are considered influential variables for the analysis, provide the evidences to support this determination!
9. What conclusions can be drawn based on the results obtained from your analysis?
10. What recommendations and/or suggestions can be offered based on the findings derived from the analysis?
11. Evaluate your best model on “Data_Prediction.csv”!

Complete each chapter to answer all the questions above. Make sure all the questions above are answered. The paper should include five chapters as follows:

Chapter 1 : Introduction

This chapter should provide background information on the topic and outline the issues discussed in the report.

Chapter 2 : Theoretical Framework

This chapter should explain the theoretical framework used in the analysis. It includes relevant theories or concepts used in the study.

Chapter 3 : Analytical Steps

This chapter should explain the steps or methods used in the analysis process from start to finish.

Chapter 4 : Analysis of Results

This chapter should contain an explanation of the findings of the analysis, overall discussion, and presentation in an easy-to-understand format.

Chapter 5 : Conclusion and Recommendation

This chapter should contain the conclusions and recommendations obtained from the analysis.