The synthetic dataset generated represents a scenario for network intrusion detection, a critical task in information security. The dataset comprises both numeric and categorical features, simulating various aspects of network traffic. Approximately 10% of the dataset contains randomly introduced missing values to mimic real-world scenarios where data may be incomplete or unavailable. This synthetic dataset is designed for training and testing network intrusion detection models, considering the complexities of real-world data, including missing values and diverse features associated with network activities. The total number of records in the dataset is 5020.

**Features overview**

1. **Numeric features**

* duration: Duration of the network activity.
* src\_bytes: Number of source bytes in the communication.
* dst\_bytes: Number of destination bytes in the communication.
* failed\_logins: Number of failed login attempts (additional numeric feature).
* num\_file\_creations: Number of file creations (additional numeric feature).

1. **Categorical features**

* protocol\_type: Type of network protocol used (e.g., tcp, udp, icmp).
* service: Type of service associated with the network activity (e.g., http, ftp, ssh).
* flag: Flag indicating the status of the network connection (e.g., SF, S0, REJ).
* logged\_in: Binary indicator of whether a user is logged in (additional categorical feature).
* is\_guest\_login: Binary indicator of whether a guest login is present (additional categorical feature).
* target: Binary indicator of network intrusion (1) and not intrusion (0).

**Possible hypothetical research questions**

These research questions aim to guide both exploratory data analysis for a deeper understanding of the dataset and data modeling to build effective intrusion detection models while considering missing values and diverse features associated with network activities.

**Data Analysis Research Questions**

1. **Exploratory data analysis (EDA)**

* How does the distribution of failed login attempts (failed\_logins) vary between instances of network intrusion and non-intrusion?
* Are there notable differences in the distribution of source bytes (src\_bytes) and destination bytes (dst\_bytes) when comparing different network protocols (protocol\_type)?
* How does the distribution of network activity duration vary across different network protocols?
* Are there discernible patterns in the distribution of source bytes and destination bytes for different network services?

1. **Missing values analysis**

* What is the impact of randomly introduced missing values on the analysis of network intrusion detection, and do certain features exhibit a higher prevalence of missing values?
* Can we identify any patterns or trends in the missing values, and do they occur more frequently in certain types of network activities or services?

**Data Modeling Research Questions**

1. **Intrusion detection modeling**

* How effectively can the provided features (duration, src\_bytes, dst\_bytes, protocol\_type, service, flag, failed\_logins, num\_file\_creations, logged\_in, is\_guest\_login) predict instances of network intrusion?
* What is the relative importance of each feature in detecting network intrusions, and are there interactions between certain features that significantly impact the model?

1. **Model generalization and robustness**

* Can the intrusion detection model generalize well across different network protocols (protocol\_type) and services (service), or do certain protocols and services require specific detection strategies?
* How robust is the predictive model when applied to instances with missing values, and are there specific features for which missing values pose a greater challenge in intrusion detection?