POLITÉCNICO DE LEIRIA | ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

COMPUTER ENGINEERING DEPARTMENT

COMPUTER ENGINEERING

ARTIFICIAL INTELLIGENCE

2023/2024 – 2nd semester

WORKSHEET X – CLASSIFICATION – TABULAR DATA

We will use the file `IoMT_Flows_Sample.csv` sample data from the IoMT dataset (available at https://zenodo.org/records/8116338). Such dataset comprises statistics on normal and malicious traffic on an Internet of Medical Things (IoMT) network. 'is_attack' differentiates normal and malicious traffic (0 = normal; 1 = malicious). The column 'traffic' identifies the normal traffic and the type of cyberattack. Your goal is to perform binary classification (i.e., identify normal and malicious traffic) and multiclass classification (i.e., identify the traffic origin) using such data. To help you complete the task, some code has already been written.

1. Open the notebook `TabularClassifier.ipynb`, look for instructions and `#TODOs`, and complete the code. The main changes you must do are:

   a) Perform one hot encoding in identified columns. Use the function *get_dummies* as in the following example:

```
df = pd.get_dummies(df, columns=['col_name'], prefix=''col_name ', dtype=int)
```

   b) Add code to normalize data in 0 to 1 range.

   c) Create training and test sets in multiclass classification.

   d) Define and compile the model for multiclass classification. Use *sparse_categorical_crossentropy* as loss function.

   e) Compute missing performance metrics for multiclass classification using macro and weighted averages and compare the results.

2. After completing and executing the code, compare the size of the files `IoMT_Flows_Sample.csv` and `IoTMT-Sample_Processed.csv`. Is there any relevant difference in size? Why?

3. Which model (binary classification or multiclass classification) achieved the best performance? Why?

Dataset P-Based Flows – Column description

- **proto** - The communication protocol used for the traffic (e.g., TCP, UDP, ICMP).
- **traffic** - Indicates the type of network activity, distinguishing between normal and potentially malicious traffic.
- **is_attack** - A binary indicator denoting whether the observed traffic is considered an attack. 0 = for normal traffic, 1 for detected attacks).
- **total_bytes** - The total amount of data transferred in bytes during the observed traffic.
- **total_pkts** - The total number of packets transmitted during the observed traffic session.
- **pkts_unidirectional_traffic** - Number of packets exchanged in unidirectional traffic.
- **pkt_difference** - The variance in packet count between different directions of traffic flow.
- **byte_difference** - The difference in data volume between different directions of traffic flow.
- **total_data_pkts** - The total number of data packets transmitted, excluding headers.
- **payload_ratio** - The ratio of payload size to the total packet size.
- **total_payload_volume** - The cumulative volume of payload data transmitted.
- **fwd_bwd_pkts_diff** - Difference in packet count between forward and backward directions.
- **duration_weighted_pkts** - Packet count weighted by duration.
- **pkts_size_weighted** - Packet count weighted by size.
- **flow_pkts_size_weighted** - Packet count within the flow weighted by size.
- **header_size_ratio** - Ratio of header size to total packet size.
- **total_header_size** - Cumulative size of headers.
- **header_size_diff** - Difference in header size between directions.
- **fwd_bwd_payload_tot_diff** - Difference in total payload size between directions.
- **fwd_bwd_payload_avg_diff** - Difference in average payload size between directions.
- **flow_fwd_payload_diff** - Difference in payload size in the forward direction within the flow.
- **flow_bwd_payload_diff** - Difference in payload size in the backward direction within the flow.
- **flow_payload_range** - Range of payload sizes within the flow.
- **iat_is_unidirectional** - Indicates if inter-arrival times are unidirectional.
- **total_activity** - Overall level of activity.
- **history_originator** - Historical activity originating from the observed source.
- **history_responder** - Historical activity corresponding to responses received by the observed source.