**Event Log for Incident Management Process**

**Dataset**

Number of instances: 141,712

Number of Features: 36

Link: <https://drive.google.com/file/d/1Qzn5vJgFQpYIqC-5o4f2x2wtGh30uywO/view?usp=sharing>

Description: This is an event log for an incident management process extracted from an audit system in Brazil. The event log is enriched with data loaded from a relational database underlying a corresponding process-aware information system. The additional information of the features is attached at the end of the document.

This project is divided into 3 deliverables, Preprocessing & EDA, Prediction & Clustering and Final Report. For all deliverables you are required to make a report that you will compile and submit in Deliverable 3 i.e., Final Report. The weightage for each deliverable is also given alongside each of them.

**D1: Preprocessing & EDA (Weightage: 30%)**

For this deliverable, focus on initial data exploration, missing data handling and graphs. This could include but not limited to:

* Removing extra information from all features in the file such as remove “Caller “, from each instance of the column, “caller\_id”. Hint, the similar update should not be in more than 13 columns in the dataset.
* Finding missing values and replacing them by an appropriate strategy (where required).
* Calculating and visualizing summary statistics.
* Finding correlation between attributes.
* Finding out dependence between categorical and numerical attributes.
* Analyzing the “incident\_state” i.e., the current state of the incident and represent these frequencies using a histogram plot.
* Finding the Average time required to handle an incident report.
* Analyzing the average number of instances reported for one incident.
* Finding the month with highest number of incidents **reported.**

**D2: Prediction and Clustering (Weightage: 40%)**

1. Clustering:

* Implementation of appropriate clustering technique on features with meaningful results. Hint. One of them may or may not include ‘closed\_code’ and ‘resolved\_by’ fields.
* Perform Outlier Analysis using appropriate fields and explain your findings.

1. Prediction:

* You are required to remove rows where priority is set to be medium and then predict the urgency and priority of the given event.
* Split your data into 80/20 ratio for training and testing data.
* Remove the Class Imbalance Ratio using K-fold Cross Validation.
* Use appropriate fields as your input features (Reminder your data must be numerical values).
* Predict the output for incident features, targeting “urgency” and “priority” using any optimizer.
* Evaluate your results on test data and provide a confusion matrix.
* Also report Precision, Recall, Accuracy and F1 score.

**D3: Final Report (Weightage: 30%)**

After each deliverable you are required to update your report. In this deliverable, focus on the following points:

* Try to apply all the concepts, learned during the course.
* You should be able to identify the approaches and techniques for each objective.
* Your conclusions, recommendations, reasoning, and findings should be supported by the relevant graphs/visualizations that best fits this incident event dataset.
* You are expected to identify what meaningful information can be extracted from this data.

**Attribute Information:**

1. number: incident identifier (24,918 values);

2. incident state: eight levels controlling the incident management process transitions from opening until closing the case;

3. active: boolean attribute that shows whether the record is active or closed/canceled;

4. reassignment\_count: number of times the incident has the group, or the support analysts changed;

5. reopen\_count: number of times the incident resolution was rejected by the caller;

6. sys\_mod\_count: number of incident updates until that moment;

7. made\_sla: boolean attribute that shows whether the incident exceeded the target SLA;

8. caller\_id: identifier of the user affected;

9. opened\_by: identifier of the user who reported the incident;

10. opened\_at: incident user opening date and time;

11. sys\_created\_by: identifier of the user who registered the incident;

12. sys\_created\_at: incident system creation date and time;

13. sys\_updated\_by: identifier of the user who updated the incident and generated the current log record;

14. sys\_updated\_at: incident system update date and time;

15. contact\_type: categorical attribute that shows by what means the incident was reported;

16. location: identifier of the location of the place affected;

17. category: first-level description of the affected service;

18. subcategory: second-level description of the affected service (related to the first level description, i.e., to category);

19. u\_symptom: description of the user perception about service availability;

20. cmdb\_ci: (confirmation item) identifier used to report the affected item (not mandatory);

21. impact: description of the impact caused by the incident (values: 1 - High; 2 - Medium; 3 - Low);

22. urgency: description of the urgency informed by the user for the incident resolution (values: 1 - High; 2 - Medium; 3 - Low);

23. priority: calculated by the system based on 'impact' and 'urgency';

24. assignment\_group: identifier of the support group in charge of the incident;

25. assigned\_to: identifier of the user in charge of the incident;

26. knowledge: boolean attribute that shows whether a knowledge base document was used to resolve the incident;

27. u\_priority\_confirmation: boolean attribute that shows whether the priority field has been double-checked;

28. notify: categorical attribute that shows whether notifications were generated for the incident;

29. problem\_id: identifier of the problem associated with the incident;

30. rfc: (request for change) identifier of the change request associated with the incident;

31. vendor: identifier of the vendor in charge of the incident;

32. caused\_by: identifier of the RFC responsible by the incident;

33. close\_code: identifier of the resolution of the incident;

34. resolved\_by: identifier of the user who resolved the incident;

35. resolved\_at: incident user resolution date and time (dependent variable);

36. closed\_at: incident user close date and time (dependent variable).