# xAutoML Course Project 2

Automating Process Discovery with AutoML for Optimal Algorithm Selection and Hyperparameter Tuning


Team 6

Institute of Computer Science, University of Tartu

# Agenda

1. Motivation and Problem Statement
2. Dataset Description
3. Project Methodology
4. Results and Discussion

# Motivation

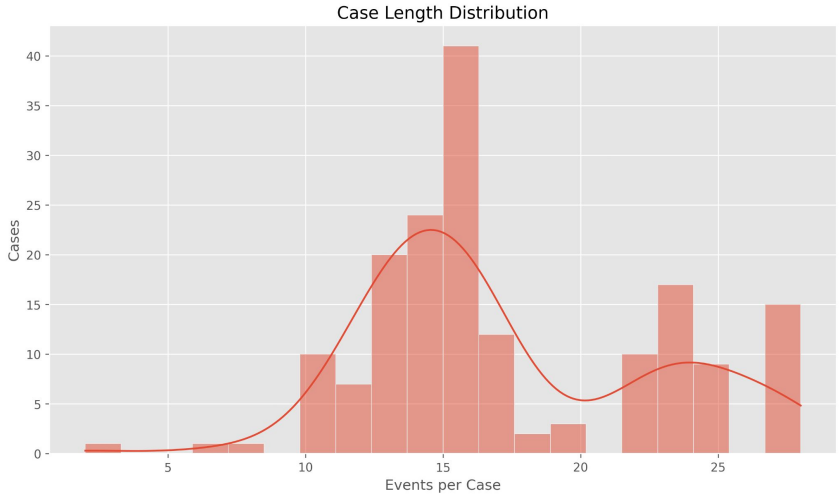
## Problem Statement

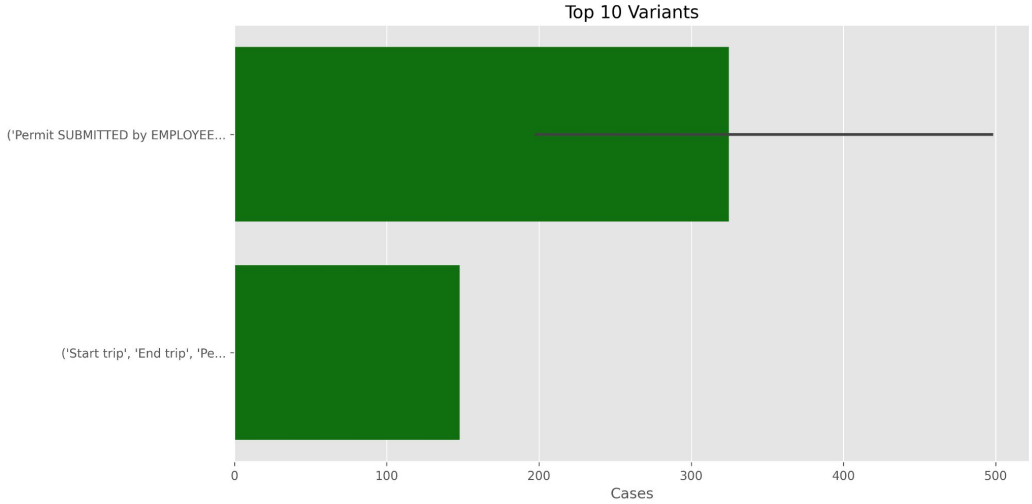
- Automated Algorithm Selection and Hyper-parameter tuning for process discovery in process mining.

- Search space is defined through 3 foundation process discovery algorithms:
  - Alpha Miner
  - Heuristic Miner
  - Inductive Miner

# Dataset Description

- Travel Permit Data BPIC Challenge 2020 including all related events of relevant prepaid travel cost declarations and travel declarations:
  - 7,065 cases,
  - 86,581 events

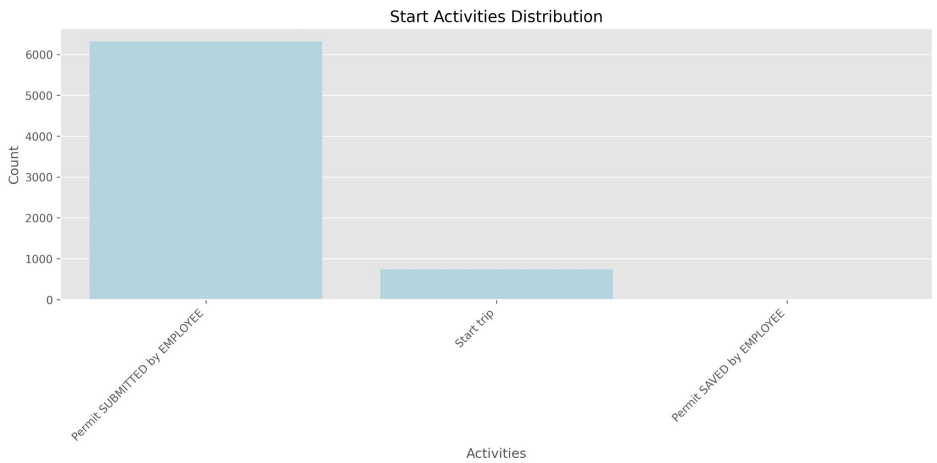


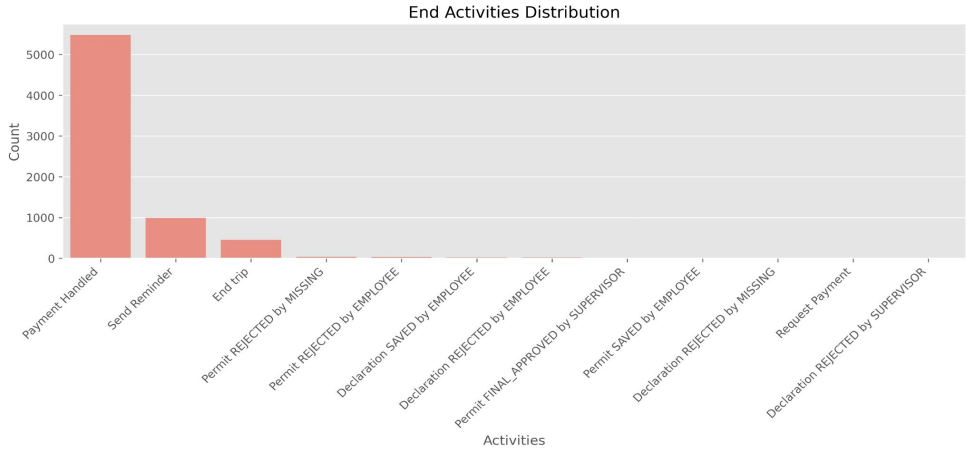



# Dataset Description

- Filtration:
  - Start Activity: Permit Submitted by Employee
  - End Activity: Payment Handled

- 6319 Traces are remaining



# Project Methodology

- Baselines with default hyper-parameters:
  - Alpha Miner
  - Heuristic Miner
  - Inductive Miner

- Experiments:
  - AutoML Frameworks:
    - Deap
    - Optuna
  - Time Budgets:
    - 5 minutes
    - 15 minutes
    - 30 minutes
    - 60 minutes
  - Search Space:
    - Alpha Miner:
      - remove_loops, ignore_noise
    - Heuristic Miner
      - dependency_thresh, noise_thresh, and_thresh
    - Inductive Miner
      - activity_freq_filter, noise_thresh, variant

- Evaluation Metrics:
  - Fitness
  - Generalization
  - Simplicity

- Statistical Test
  - Friedman Nemenyi Test

# Project Methodology

- Baselines with default hyper-parameters:
  - Alpha Miner
  - Heuristic Miner
  - Inductive Miner

| Approach | Fitness | Generalization | Simplicity |
|---|---|---|---|
| Alpha Miner | 0.4199 | 0.419 | 0.8553 |
| Heuristic Miner | 0.8892 | 0.4768 | 0.6057 |
| **Inductive Miner** | **1** | **0.5947** | **0.8617** |

- Evaluation Metrics:
  - Fitness
  - Generalization
  - Simplicity

# Project Methodology

| Objective | Time Budget | Approach | Fitness | Generalization | Simplicity | Best Algorithm |
|---|---|---|---|---|---|---|
| Fitness | **5** | **Deap/HyperOpt** | **1** | **0.9857** | **1** | **Inductive Miner** |
| | | **Optuna** | **0.883** | **0.5736** | **0.4901** | **Heuristic Miner** |
| | 15 | Deap/HyperOpt | 1 | 0.9857 | 1 | Inductive Miner |
| | | Optuna | 1 | 0.9857 | 1 | |
| | 30 | Deap/HyperOpt | 1 | 0.9857 | 1 | |
| | | Optuna | 1 | 0.9857 | 1 | |
| | 60 | Deap/HyperOpt | 1 | 0.9857 | 1 | |
| | | Optuna | 1 | 0.9857 | 1 | |

| Approach | Fitness | Generalization | Simplicity |
|---|---|---|---|
| Alpha Miner | 0.4199 | 0.419 | 0.8553 |
| Heuristic Miner | 0.8892 | 0.4768 | 0.6057 |
| **Inductive Miner** | **1** | **0.5947** | **0.8617** |

- Experiments:
  - AutoML Frameworks:
    - Deap
    - Optuna
  - Time Budgets:
    - 5,15,30,60 minutes
  - Search Space:
    - Alpha Miner:
      - remove_loops, ignore_noise
    - Heuristic Miner
      - dependency_thresh, noise_thresh, and_thresh
    - Inductive Miner
      - activity_freq_filter, noise_thresh, variant
- Evaluation Metrics:
  - Fitness
  - Generalization
  - Simplicity
- Statistical Test
  - Friedman Nemenyi Test

# Project Methodology

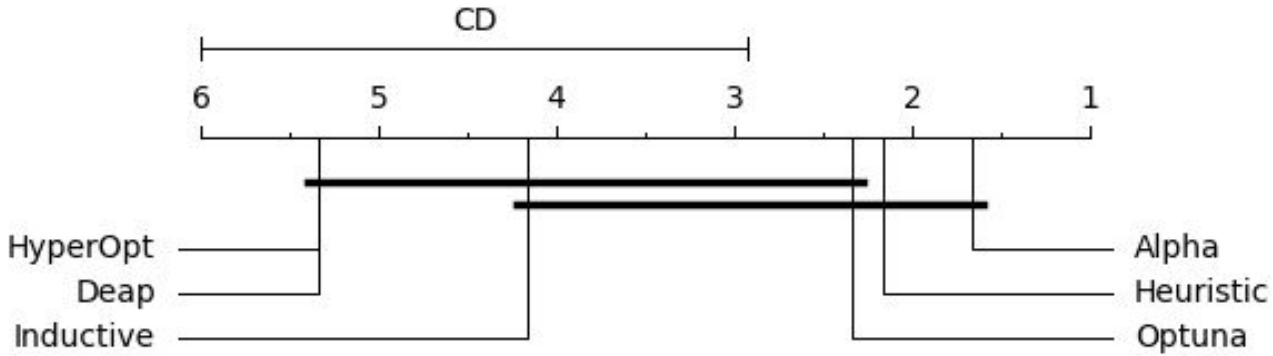
| Objective | Time Budget | Approach | Fitness | Generalization | Simplicity | Best Algorithm |
|---|---|---|---|---|---|---|
| Fitness +Generalization+Simplicity | 5 | **Deap** | **1** | **0.9857** | **1** | **Inductive Miner** |
| | | **Hyperopt** | **1** | **0.9857** | **1** | |
| | | **Optuna** | **0.906** | **0.7131** | **0.5356** | **Heuristic Miner** |
| | 15 | Deap/Optuna/hyperopt | 1 | 0.9857 | 1 | Inductive Miner |
| | 30 | Deap/Optuna/hyperopt | 1 | 0.9857 | 1 | |
| | 60 | Deap/ Optuna/ hyperopt | 1 | 0.9857 | 1 | |

| Approach | Fitness | Generalization | Simplicity |
|---|---|---|---|
| Alpha Miner | 0.4199 | 0.419 | 0.8553 |
| Heuristic Miner | 0.8892 | 0.4768 | 0.6057 |
| **Inductive Miner** | **1** | **0.5947** | **0.8617** |

- Experiments:
  - AutoML Frameworks:
    - Deap
    - Optuna
  - Time Budgets:
    - 5,15,30,60 minutes
  - Search Space:
    - Alpha Miner:
      - remove_loops, ignore_noise
    - Heuristic Miner
      - dependency_thresh, noise_thresh, and_thresh
    - Inductive Miner
      - activity_freq_filter, noise_thresh, variant
- Evaluation Metrics:
  - Fitness
  - Generalization
  - Simplicity
- Statistical Test
  - Friedman Nemenyi Test

# Project Methodology

- Friedman Nemenyi Test:
  - Defaults Vs HPO Frameworks at 5 Min time budget

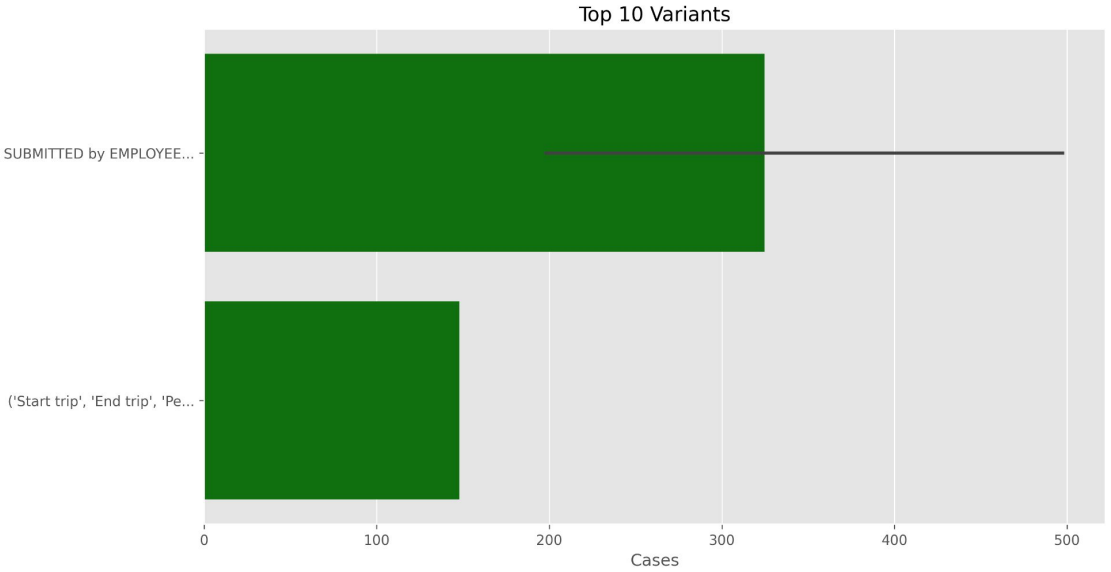


# Conclusions

- Search space for process discovery is not huge like machine learning.
- Hyper-parameter optimization does not enhance the models performance.
- Most of the hyper-parameters does not affect the performance.
- Algorithm Selection step is the most important stage (at least for all BPIC datasets).
  - Can be solved by some defined rules
  - Can be solved with Meta-Learning and algorithm selection Only
- The dataset selection was not a good one for this task as we do not have variants for the same process.

# Conclusions

- Search space for process discovery is not huge like machine learning.
- Hyper-parameter optimization does not enhance the models performance.
- Most of th
- Algorithm
  - Can b
  - Can b
- The datas




e.
or all BPIC
datasets).
Only
not have
variants fo

# Thanks for your attention!

Team Members:

1. Ahmed Wael
2. Noel Bosch
3. Mohamed Maher

Find more about our work:

Source Code: 

# Project Items

1. Dataset Exploration Preparation and Cleansing → A.Wael
    a. BPIC
    b. HelpDesk
2. Algorithms:
    a. Alpha Miner
    b. Heuristic Miner
    c. Inductive Miner
3. Baselines Benchmarking (Algorithms with Default Hyper-parameters) → A.Wael (Eo Wednesday)
4. Definition of the Configuration Space → Maher, Noel
5. Auto CASH for
    a. Approaches:
        i. Optuna → Maher (Eo Saturday)
        ii. Hyper-Opt → Noel (Eo Saturday)
    b. Time Budgets:
        i. 15 min,
        ii. 30min,
        iii. 60min
    c. Metrics:
        i. Multi-objective (Fitness, simplicity and generalization) 1 + 1 + 1
        ii. Single objective (Fitness)
6. Statistical Test (60 min) → Noel (Eo Sunday)
7. Presentation Preparation → Maher (Sunday)