Meta Feature classification and insertion on a Synthetic dataset generation process

João Carlos Fonseca Pina de Lemos - 201000660

Supervisor : Daniel Castro Silva
Co-Supervisor: Leonardo Ferreira

July 26, 2022

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

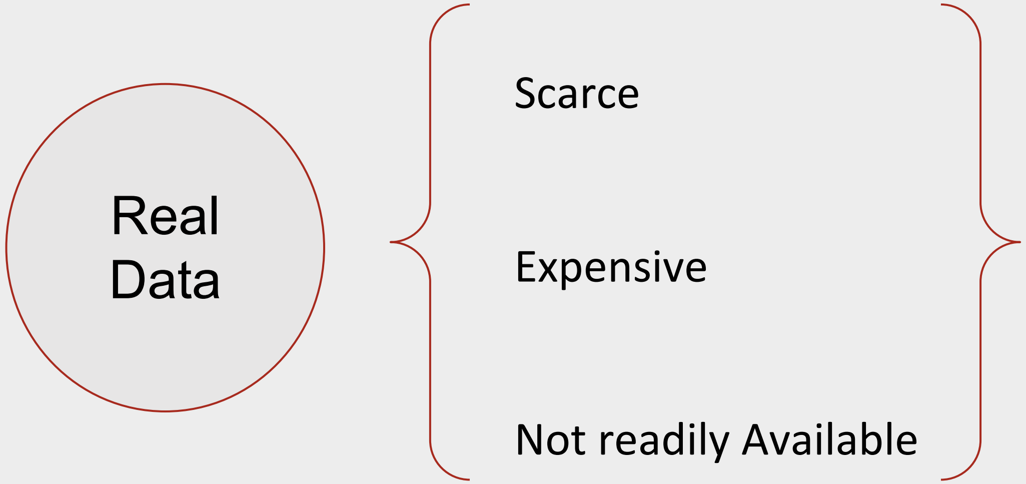
Master in Informatics and Computing Engineering

0. Structure

- Introduction
- Goals
- SNOOKER – An Overview
- Meta feature extraction
- Meta feature insertion
- Conclusion
- Future Works

1. Introduction

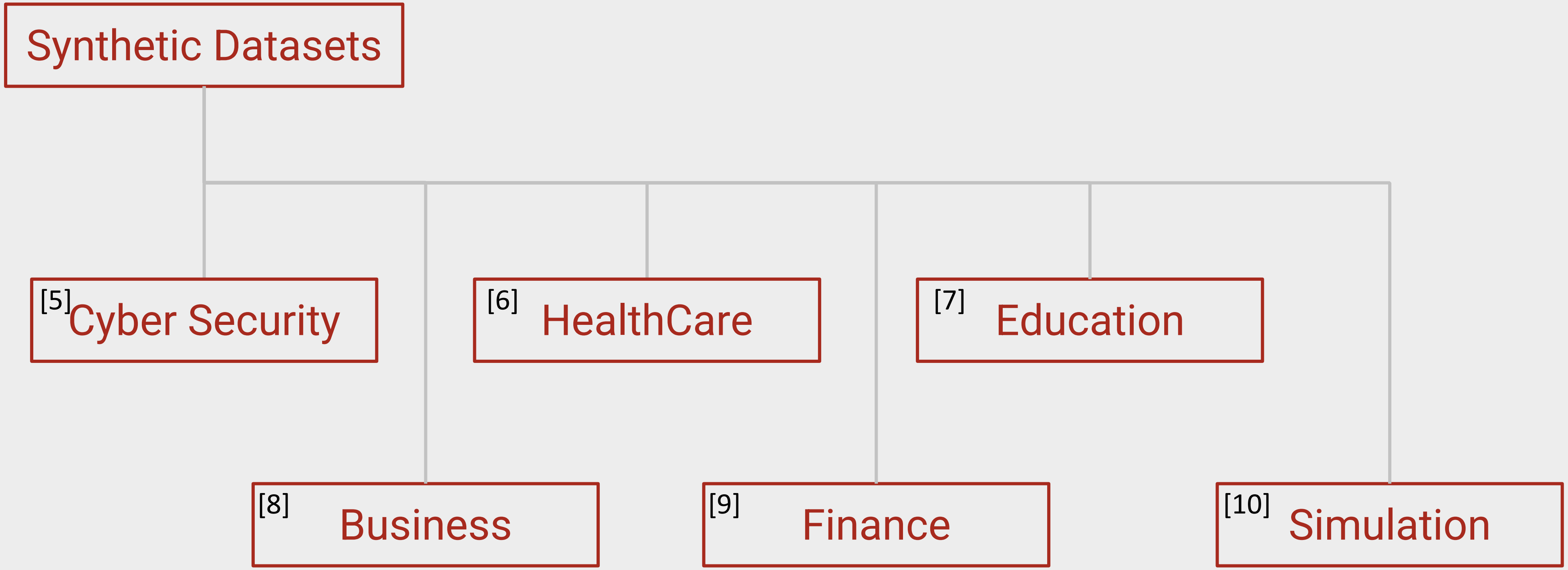
Dataset is a term used to describe a collection of data. A small dataset is also insufficient to produce an accurate prediction model.



Synthetic datasets are excellent alternatives [2].

[2] Anantrasirichai, N., Biggs, J., Albino, F., and Bull, D. (2019). A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets. Remote Sensing of Environment, 230:111179

1. Introduction

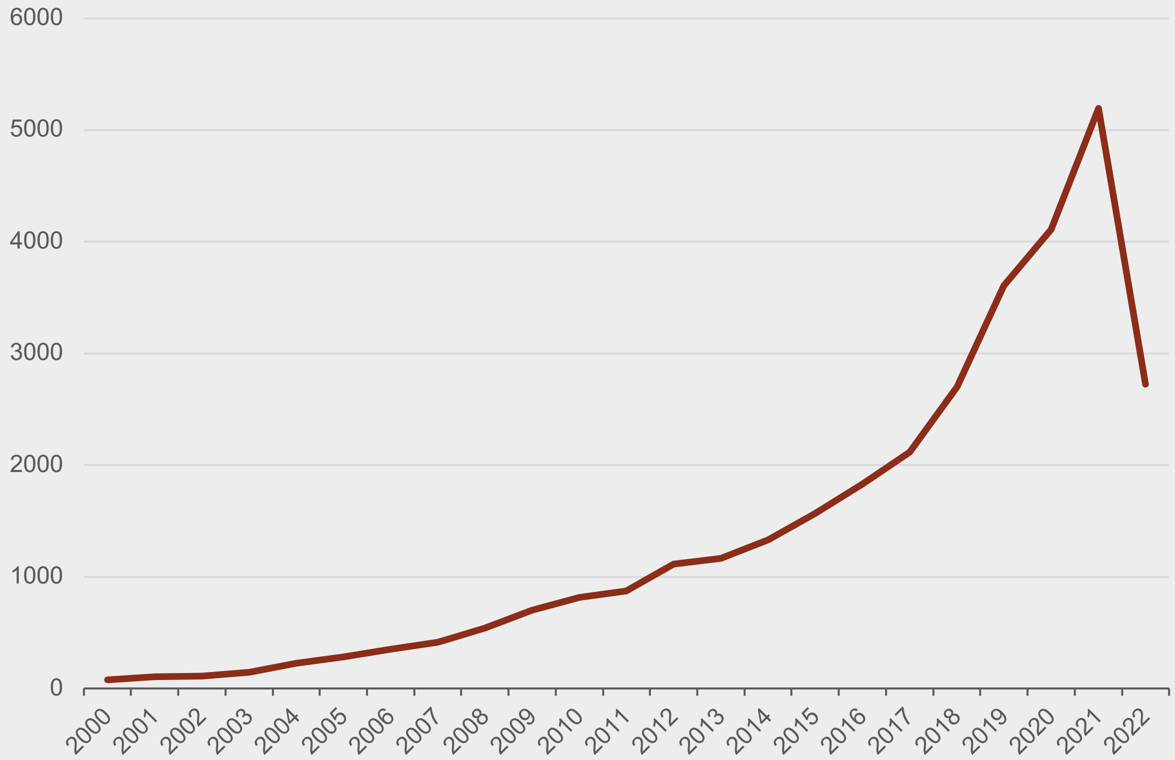


[5] - O'Shaughnessy, S., & Gray, G. (2011). Development and evaluation of a dataset generator tool for generating synthetic log files containing computer attack signatures. International Journal of Ambient Computing and Intelligence (IJACI), 3(2), 64-76.
[6] - Dahmen, J., & Cook, D. (2019). SynSys: A synthetic data generation system for healthcare applications. Sensors, 19(5), 1181.
[7] - Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review, 96, 346-353.
[8] - Mehdiyev, N., Evermann, J., & Fettke, P. (2017, July). A multi-stage deep learning approach for business process event prediction. In 2017 IEEE 19th conference on business informatics (CBI) (Vol. 1, pp. 119-128). IEEE.
[9] - Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K. (2020). Intelligent hybrid model for financial crisis prediction using machine learning techniques. Information Systems and e-Business Management, 18(4), 617-645.
[10] - Kar, A., Prakash, A., Liu, M. Y., Cameracci, E., Yuan, J., Rusiniak, M., ... & Fidler, S. (2019). Meta-sim: Learning to generate synthetic datasets. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4551-4560).

1. Introduction

- Synthetic datasets are a valid option to use on Machine Learning models

- Synthetic datasets are a growing subject with a lot left to explore



Papers on synthetic datasets in Scopus from 2000 to July 2022

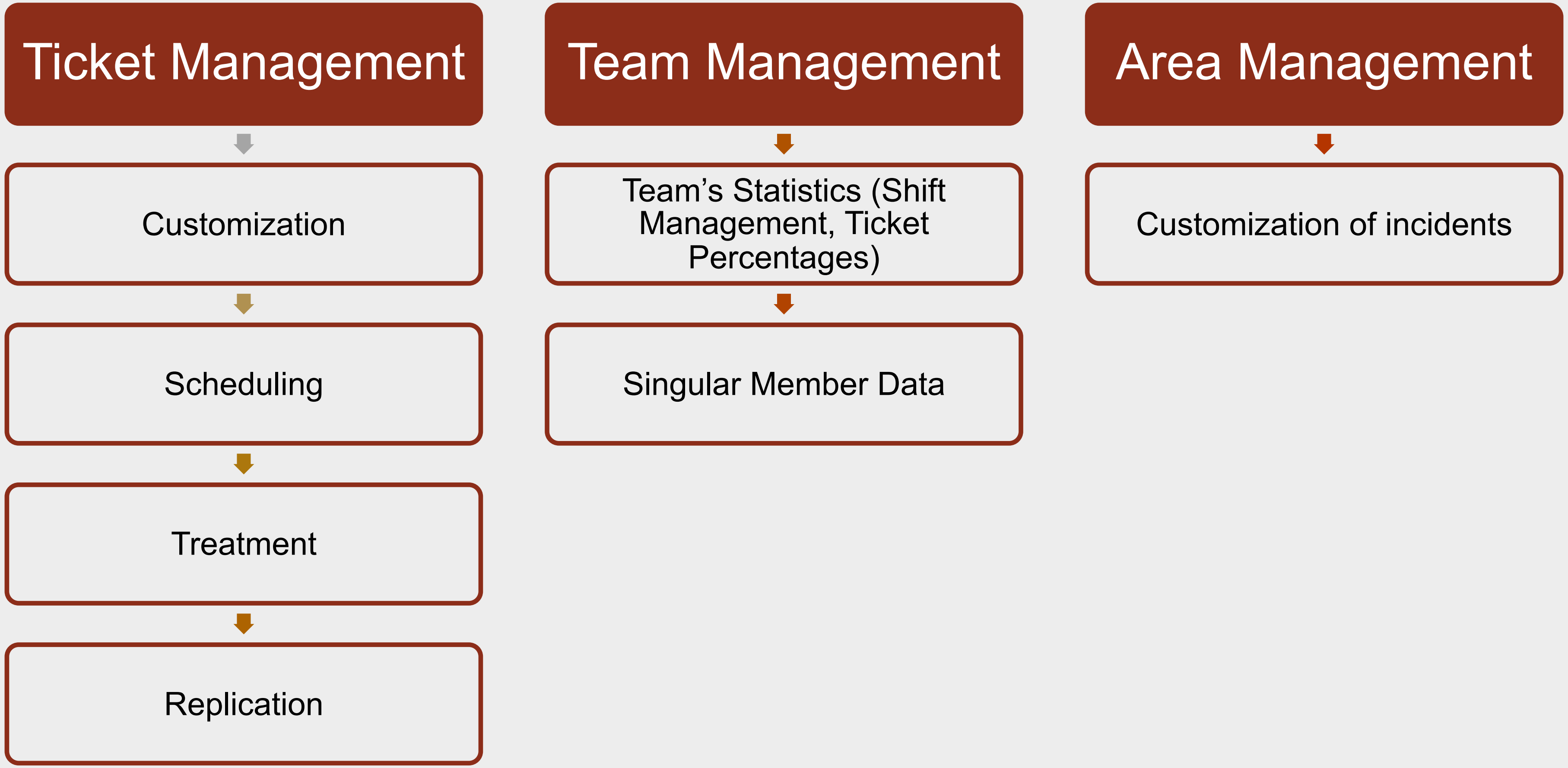
2. Goals

- Main Goal: Implementation of Meta Features Extraction and study of inclusion on a Synthetic Dataset Generation Process
    - Analysis, selection and compilation of meta features list
    - Implementation of meta feature extraction tool to the current generator
    - Analysis of methods for inserting meta features in the generation process

3. SNOOKER – An Overview

- SNOOKER: A DataSet GeNeratOr fOr HelpdesK SERvices
- Developed by Leonardo Ferreira (Co-Supervisor)
- Generates realistic ticket-based datasets
- Uses information provided by the user and by a YAML configuration file

3. SNOOKER – An Overview



Ticket Management

Customization

Scheduling

Treatment

Replication

Team Management

Team’s Statistics (Shift Management, Ticket Percentages)

Singular Member Data

Area Management

Customization of incidents

3. SNOOKER – An Overview

Preliminary Data

- Id
- Location
- Raised (UTC)
- Client
- Family & Subfamily
- Actions
- Escalation
- Outliers

Treatment Data

- Allocated and Fixed
- Stages
- Team
- Users Available
- Users Actions
- User and Actions Chosen
- Action Chosen Status
- Action Chosen Duration
- User Shift
- Similar Tickets
- Inheritance Elapsed Time
- Status

Domain Extra Data

- Coordinated Tickets
- Destination & Source IPs
- Destination and Source Ports

4. Meta Feature Extraction

Datasets are collections of data; as data differs, datasets also differ, having different characteristics. These characteristics are called meta-features and work as the fingerprint of a dataset [3]

[3] - Reif, M., Shafait, F., Goldstein, M., Breuel, T., and Dengel, A. (2014). Automatic classifier selection for non-experts. Pattern Analysis and Applications, 17(1):83–96.

4. Meta Feature Extraction

General

- Directly Observable
- Describe basic Information
- Most Straightforward

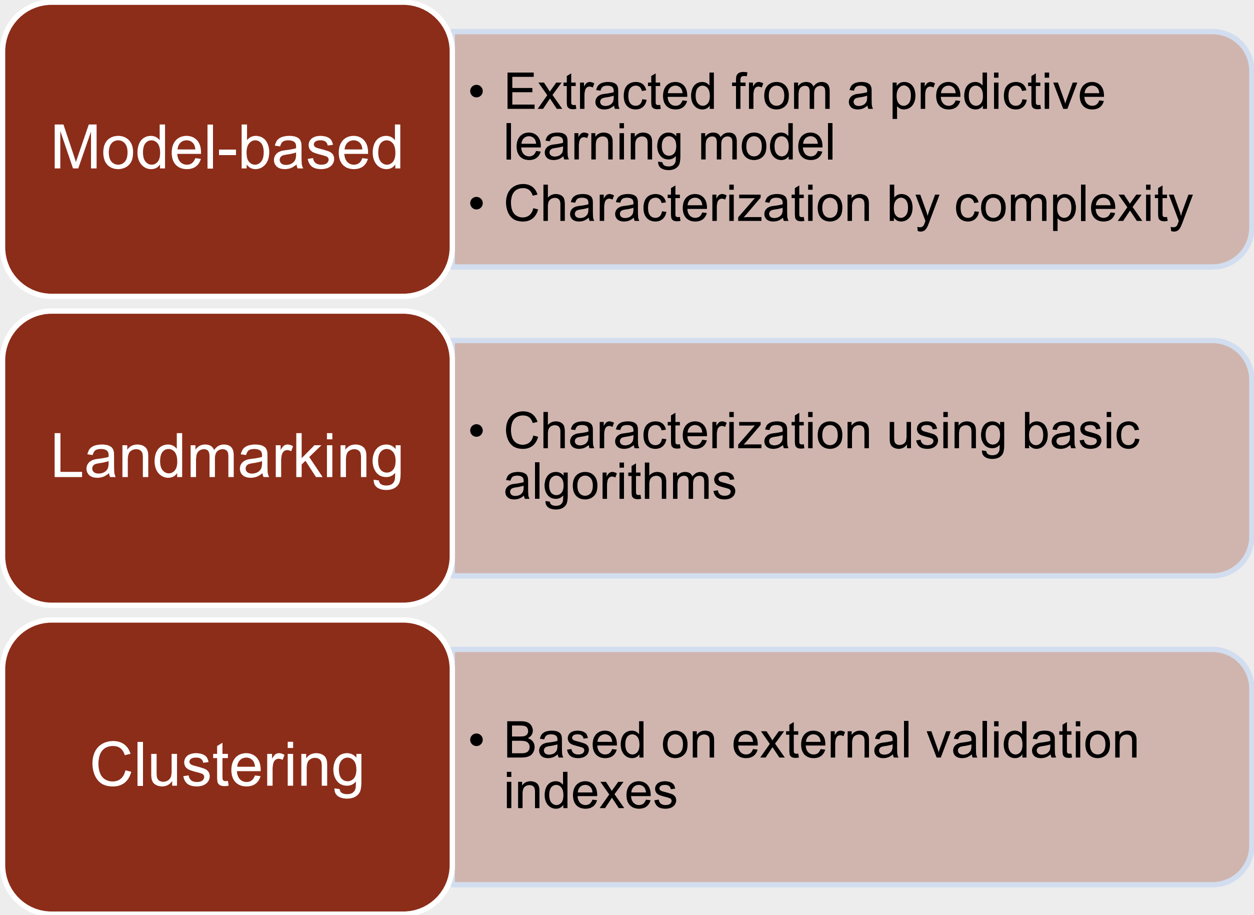
Statistical

- Reflect performance of statistical algorithms
- Largest and most diverse group

Information-Theoretic

- Reflect the amount of information in the data
- Restricted to classification problems

4. Meta Feature Extraction



Model-based

- Extracted from a predictive learning model
- Characterization by complexity

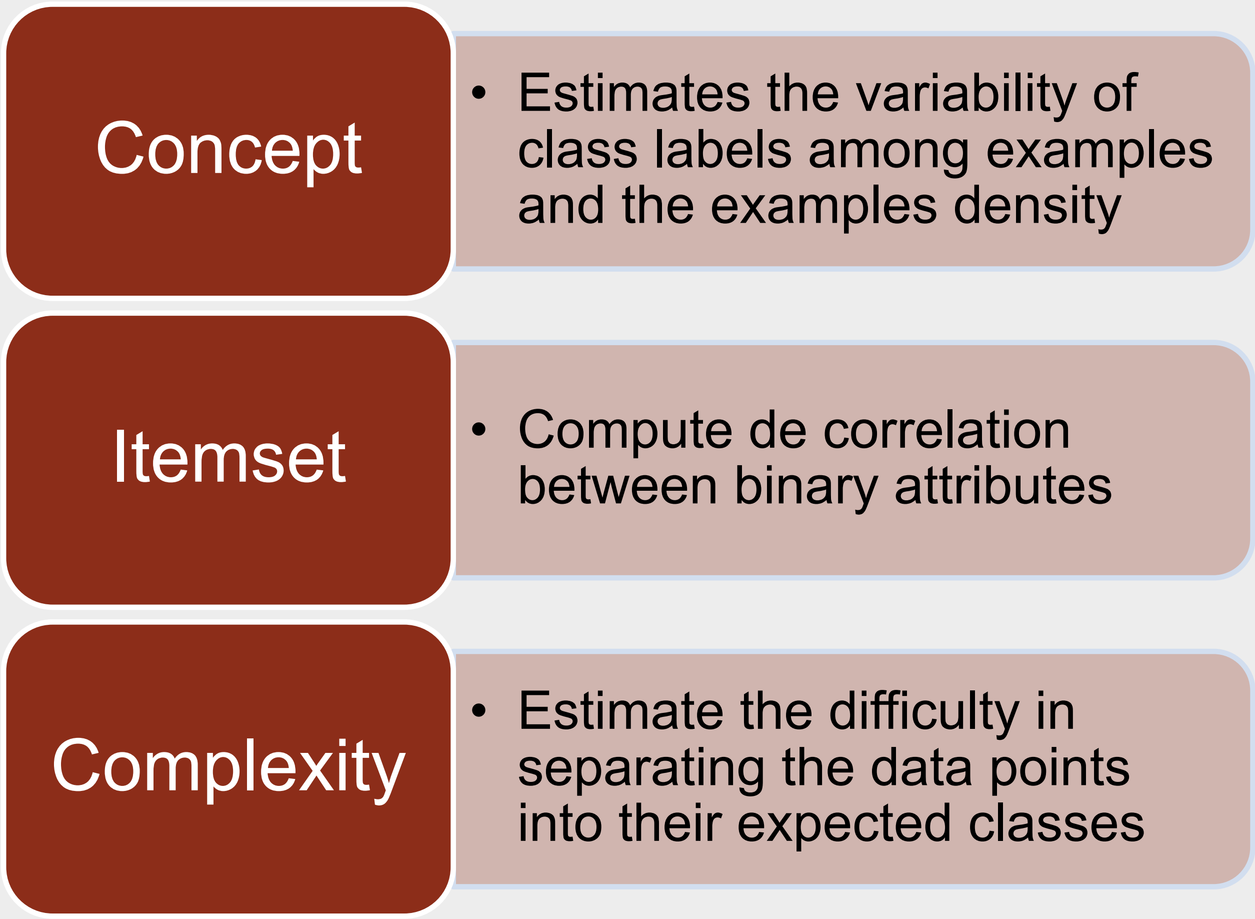
Landmarking

- Characterization using basic algorithms

Clustering

- Based on external validation indexes

4. Meta Feature Extraction



Concept

- Estimates the variability of class labels among examples and the examples density

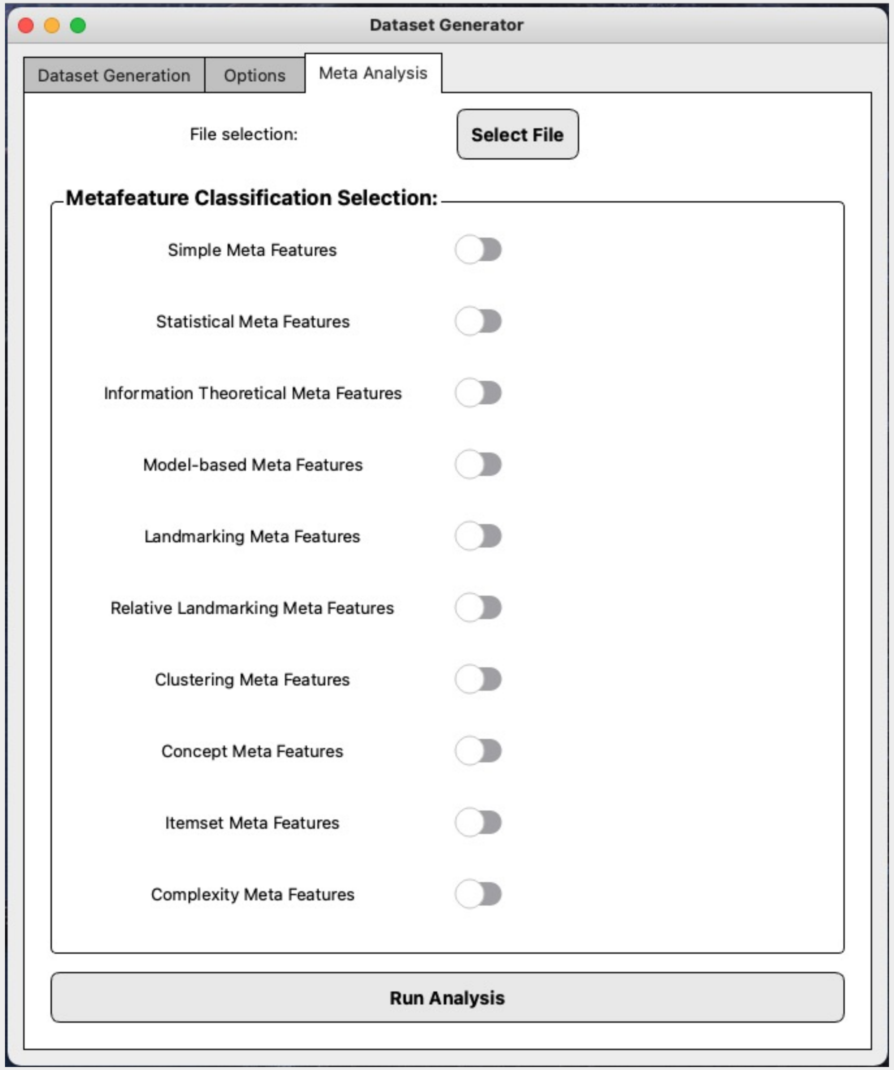
Itemset

- Compute de correlation between binary attributes

Complexity

- Estimate the difficulty in separating the data points into their expected classes

4. Meta Feature Extraction



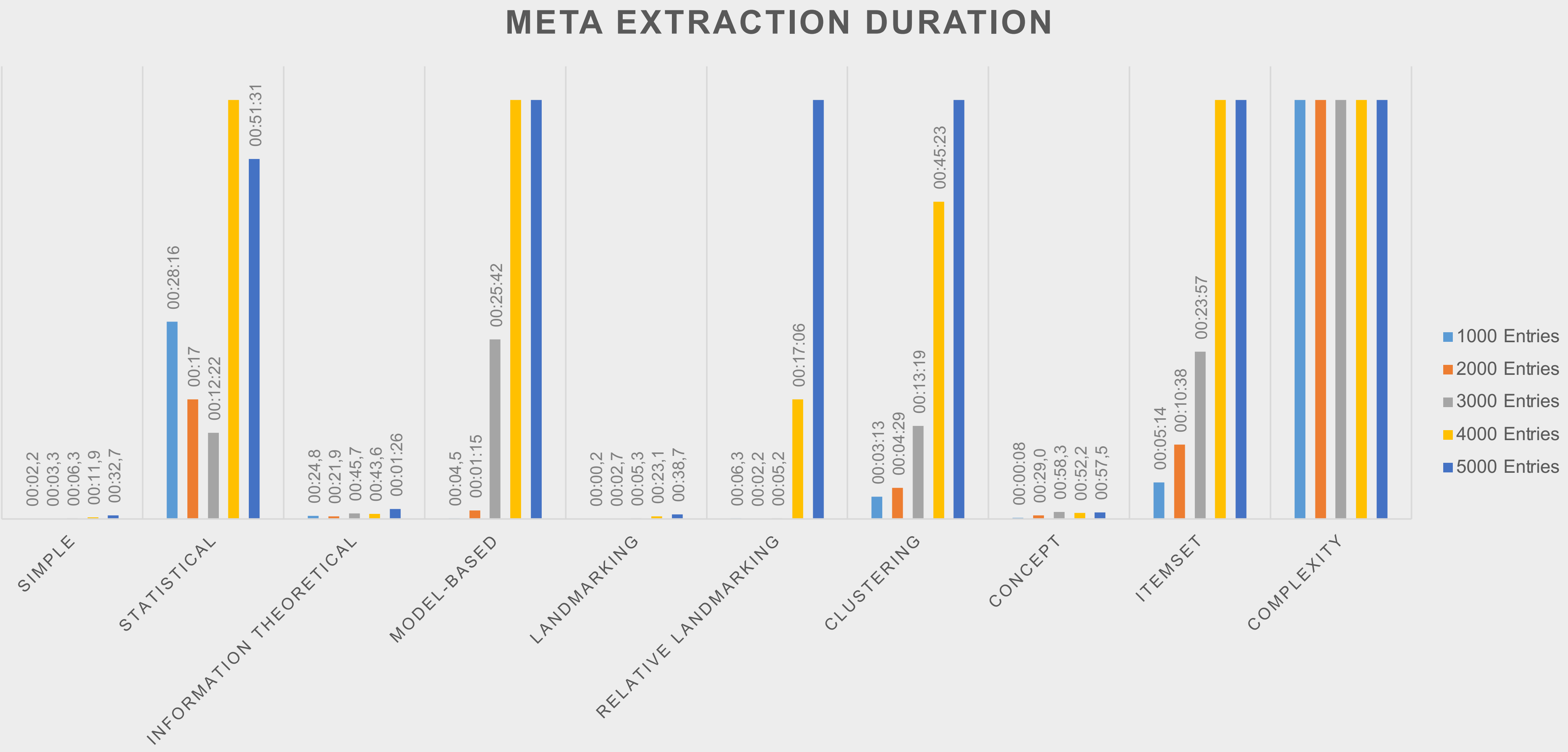
File selection:

Metafeature Classification Selection:

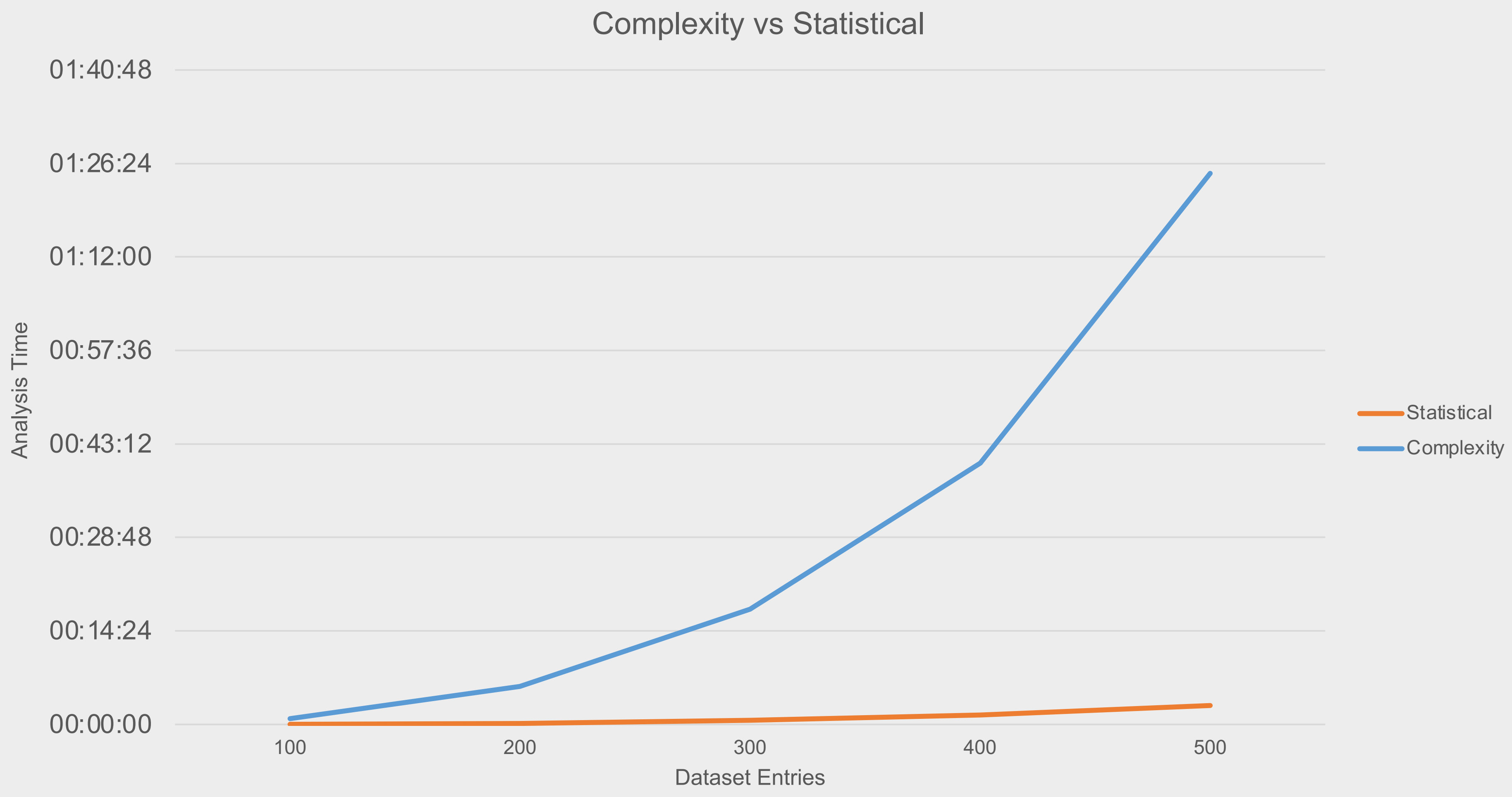
Simple Meta Features ○
Statistical Meta Features ○
Information Theoretical Meta Features ○
Model-based Meta Features ○
Landmarking Meta Features ○
Relative Landmarking Meta Features ○
Clustering Meta Features ○
Concept Meta Features ○
Itemset Meta Features ○
Complexity Meta Features ○

- Started by making extraction from scratch
- Slow implementation lead us to use an external library to make calculations - pymfe
- User interface developed inside SNOOKER
- Users can select a .CSV file and pick meta feature families to extract
- Extraction achieved good results
- Extraction was <u>time-consuming</u>

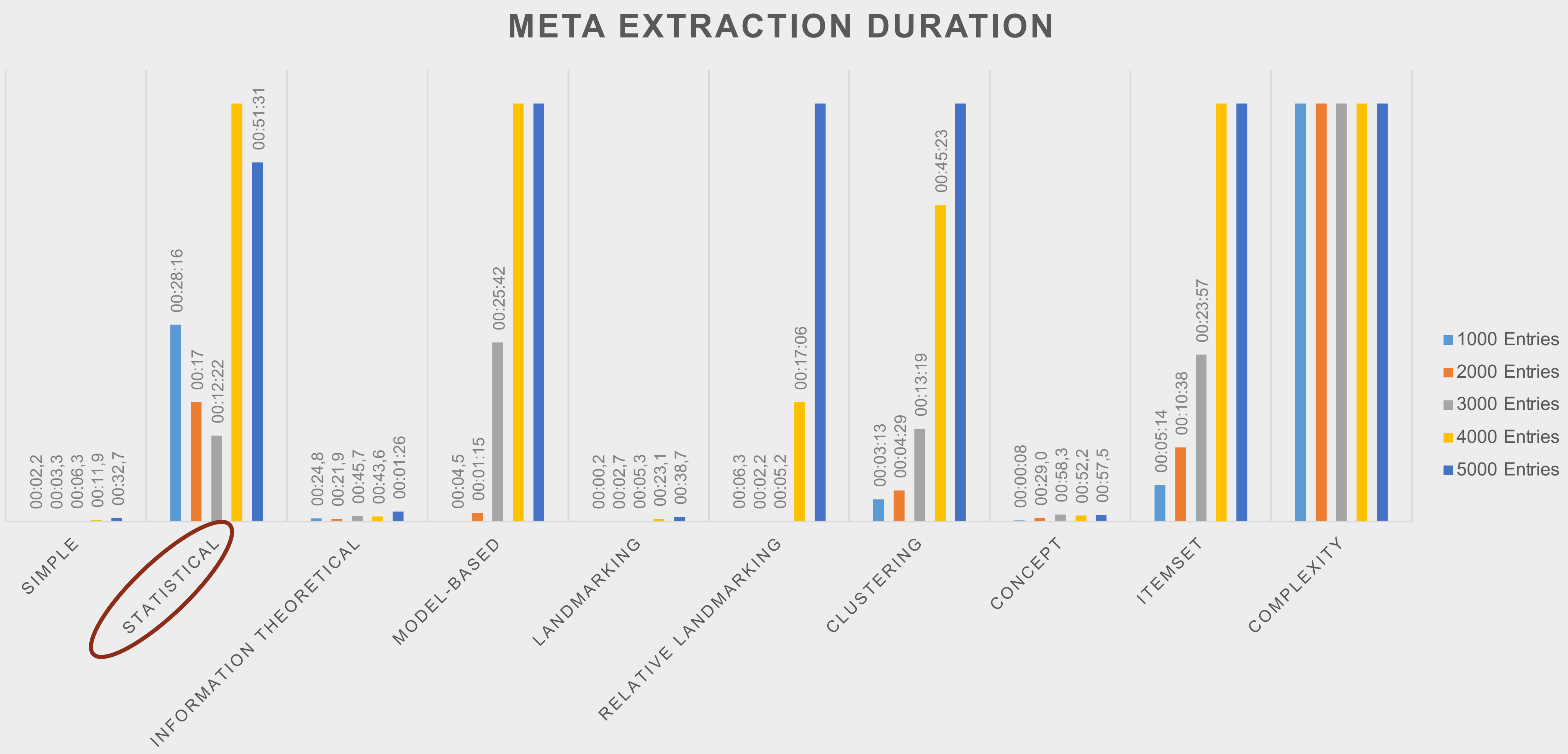
4. Meta Feature Extraction



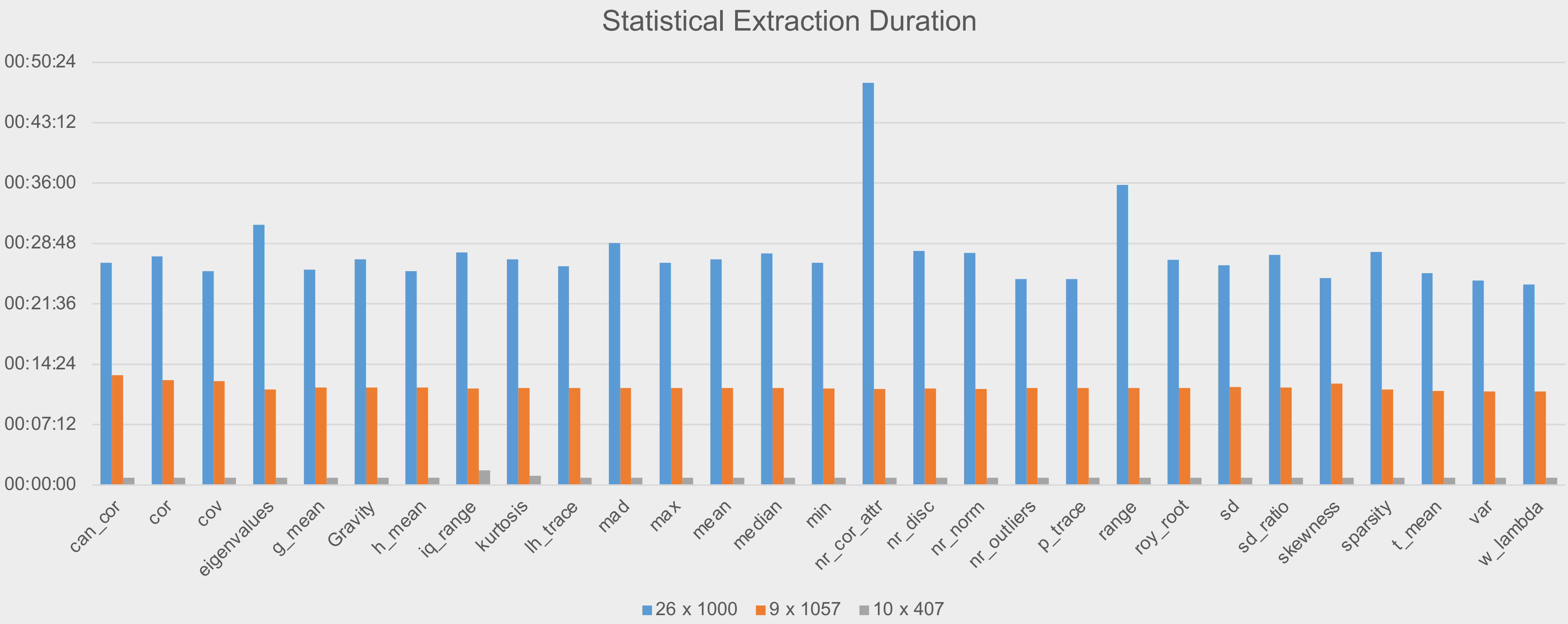
4. Meta Feature Extraction



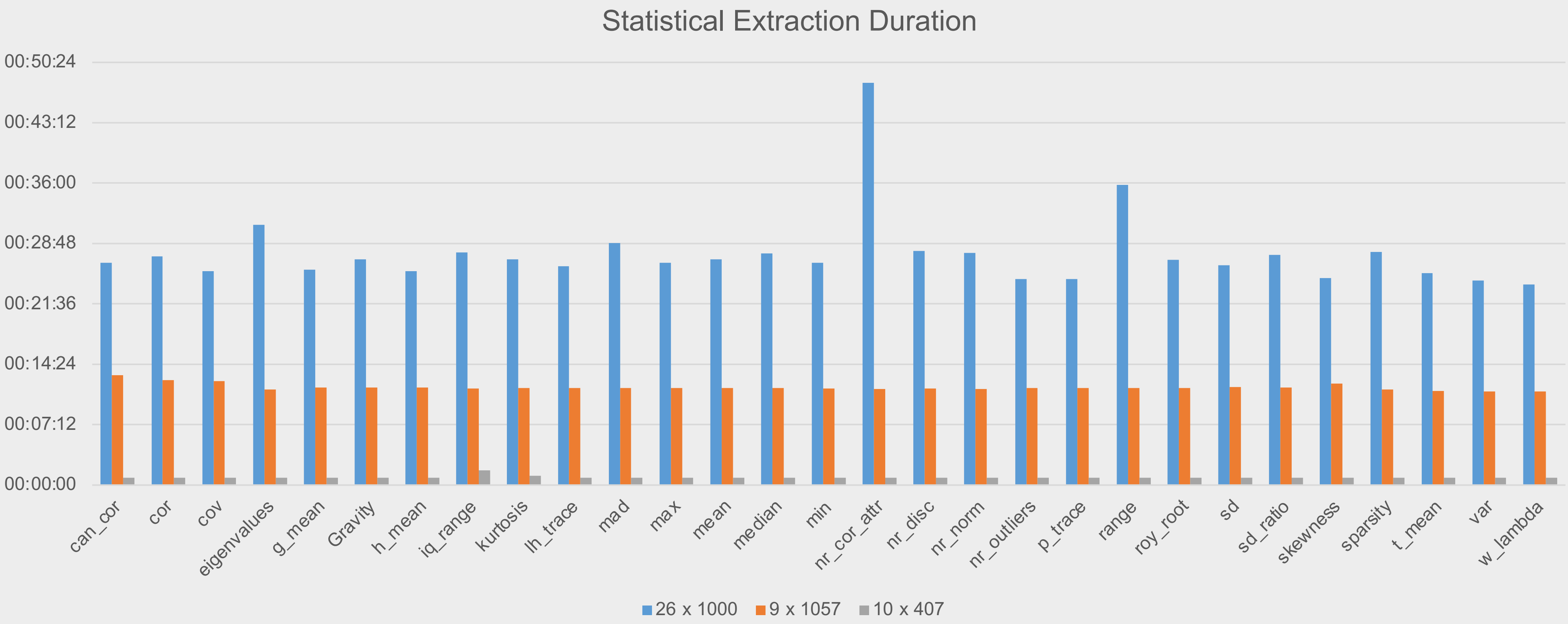
4. Meta Feature Extraction



4. Meta Feature Extraction



4. Meta Feature Extraction



Despite some peaks, no real outlier was found

4. Meta Feature Extraction

We ended up with a fully functional extractor tool that can extract a large amount of meta features form a dataset.

Meta-extraction tasks can take a lot of time to compute.

Meta-extraction requires a good amount of computational power.

5. Meta Feature Insertion

- Synthetic datasets with specific meta features can help generate better datasets.
- Randomly generated datasets cannot ensure a good distribution of meta features in those datasets.
- Generating datasets with specific values of meta features allows more controlled experiments.

1. Can we force meta features onto the generation process?
2. Would this insertion severely impact the rest of the dataset?

5. Meta Feature Insertion

- Conditional Distribution
- Bayesian Network
- Mixture of product of Multinomials
- Generative Adversarial Networks
- Linear Regression
- Decision Tree
- Random Forecast

5. Meta Feature Insertion

Conditional Distribution

Bayesian Network

Mixture of product of Multinomials

Generative Adversarial Networks

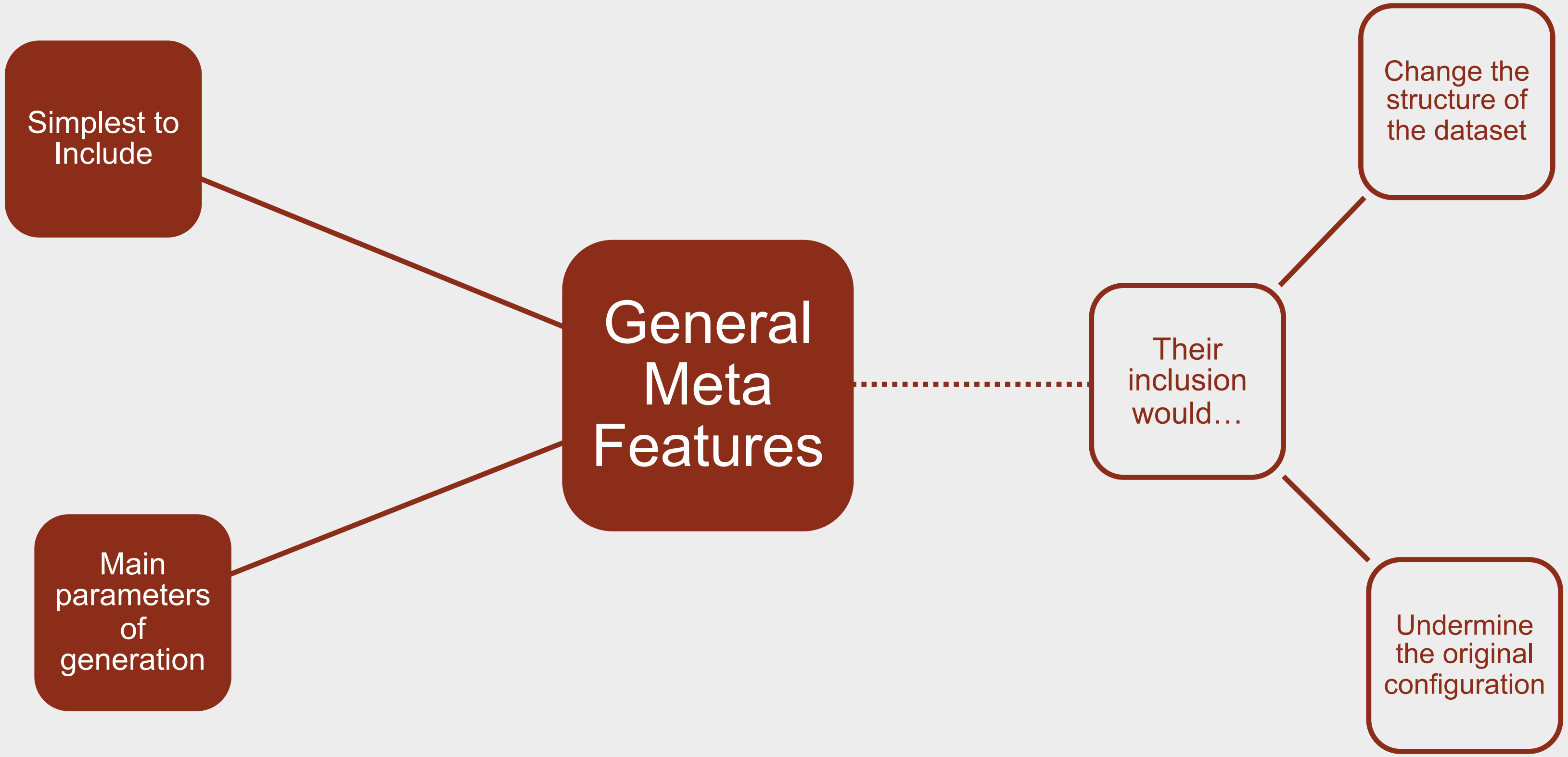
Linear Regression

Decision Tree

Random Forecast

5. Meta Feature Insertion

General Meta Features:



General Meta Features

Simplest to Include

Main parameters of generation

Their inclusion would…

Change the structure of the dataset

Undermine the original configuration

5. Meta Feature Insertion

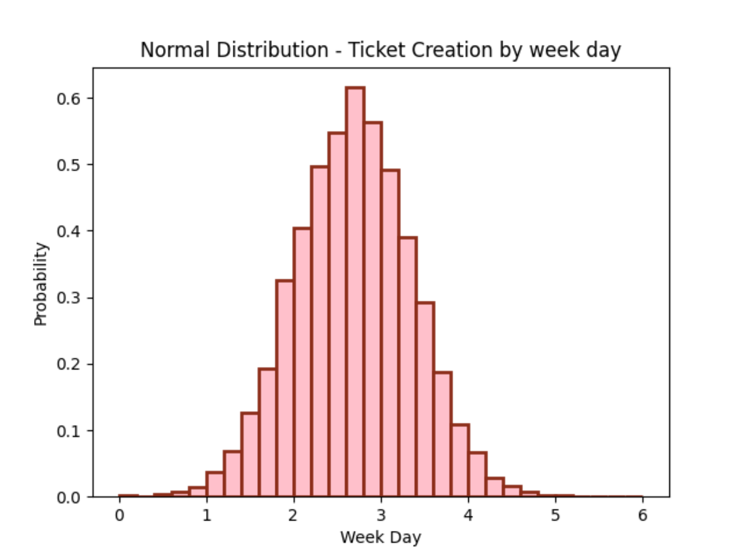
Statistical Meta Features:

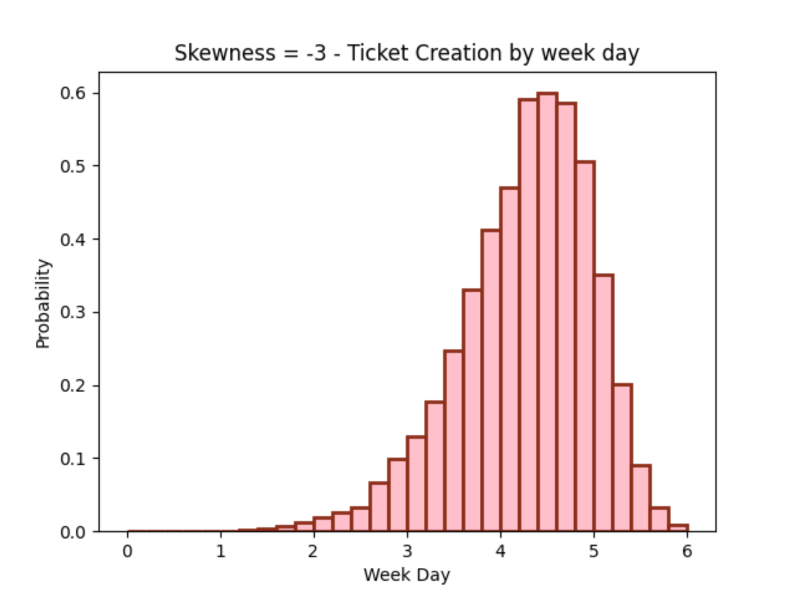
Would facilitate the creation of more diverse and valid datasets

Datasets with different characteristics

Meta insertion applied in runtime

Applied on singular or small groups of columns per case





Ex.: Inserting Skewness of value -3 into ticket generation by day of the week.

5. Meta Feature Insertion

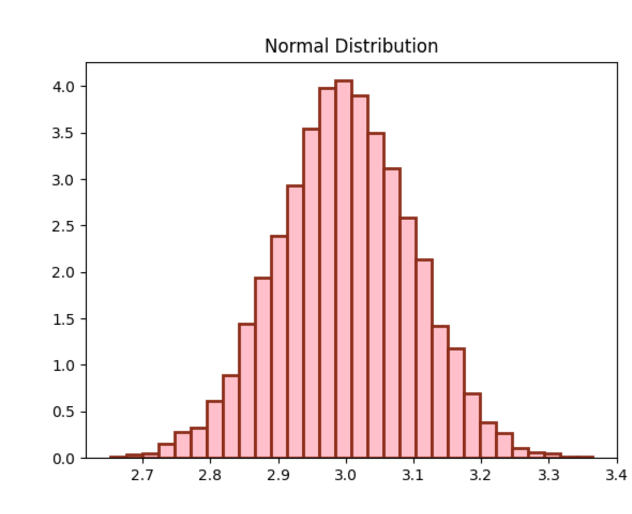
Statistical Meta Features:

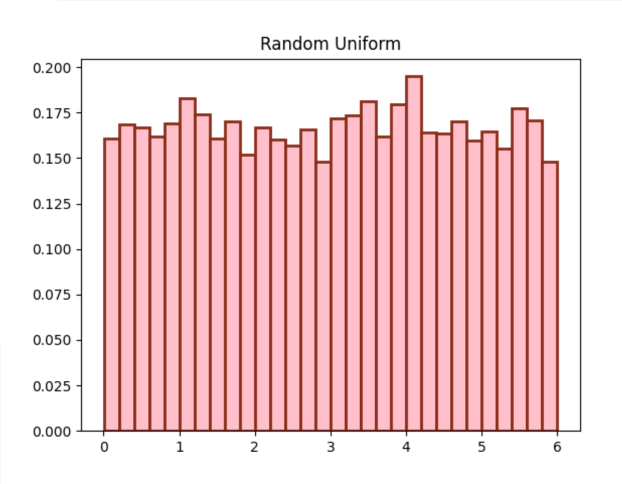
Would facilitate the creation of more diverse and valid datasets

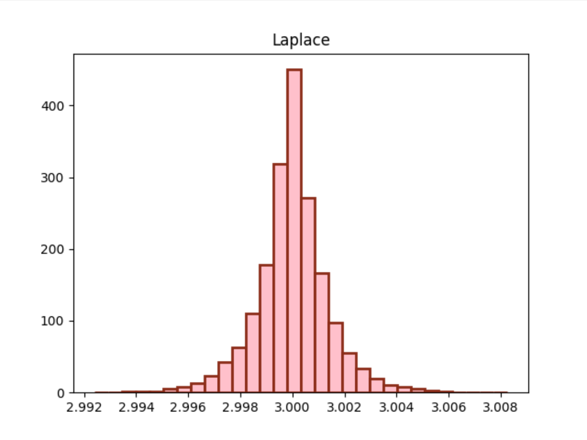
Datasets with different characteristics

Meta insertion applied in runtime

Applied on singular or small groups of columns per case



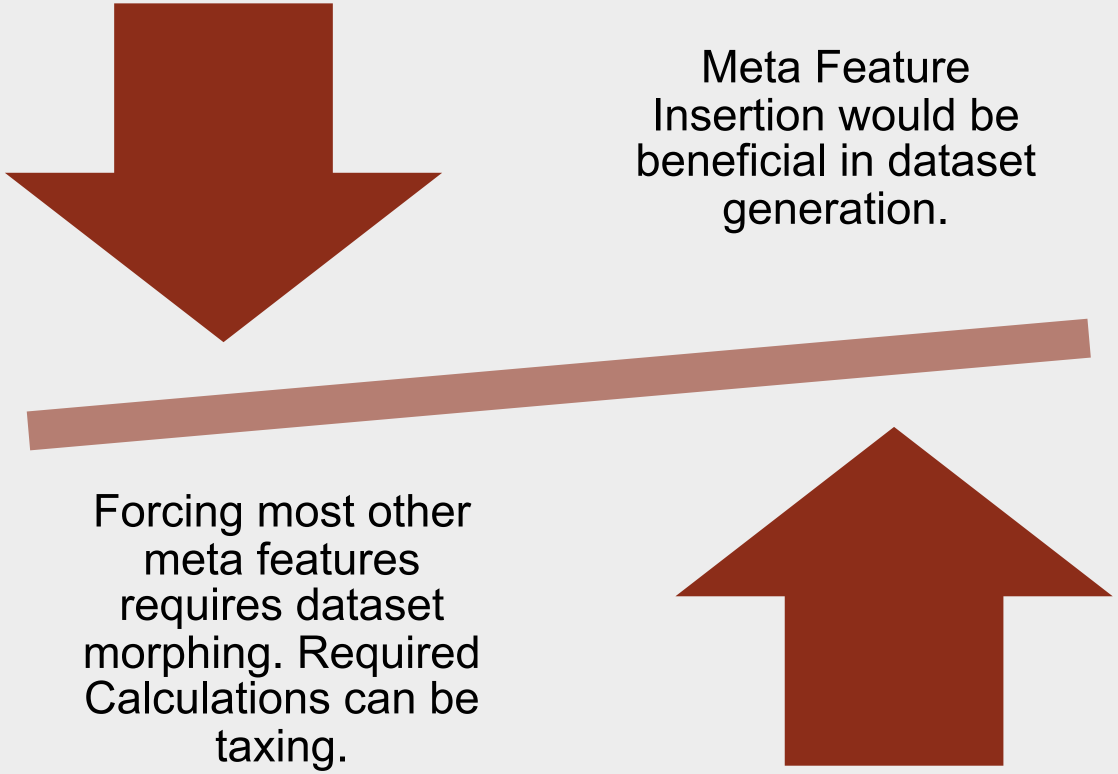




Ex.: Kurtosis influence on ticket generation by day of the week. Normal distribution has as Gaussian kurtosis of 2.96, Random Uniform has a kurtosis value of 1.8 and Laplace distribution has a kurtosis of 6.45.

5. Meta Feature Insertion

Other Meta Features:



Meta Feature Insertion would be beneficial in dataset generation.

Forcing most other meta features requires dataset morphing. Required Calculations can be taxing.

5. Meta Feature Insertion

Impact of Meta-Insertion:

During Generation

- □ Can guarantee the stability of the dataset
- □ Other meta features values will change

Dataset Morphing

- □ Requires freezing of generation processes past the ones impacted by the insertion
- □ More time and resource consuming
- □ Cannot guarantee stability of dataset

Meta-integration should be looked at on a column-by-column basis first, looking at the impact that change will have on the rest of the generation later. Especially on generation processes with synthetic dataset outputs of higher dimensionality.

6. Conclusions

We developed a classification tool using the pyfme library to help us extract as many meta features from generated datasets as possible, allowing for vast classification options.

Inclusion of meta features on a generation process is a complicated subject. While inclusion is possible should be looked at with care, spetially on bigger datasets.

7. Future Works

After concluding this study here are the possible paths to evolve it:

1. Transform SNOOKER into a web service.
2. Improve meta-extractor tool to allow for more calculations (ex.: allow missing values) and more file types. Imporvements to UI and UX.
3. More POC studies on meta-inclusion and eventual feature development on SNOOKER.

Meta Feature classification and insertion on a Synthetic dataset generation process

João Carlos Fonseca Pina de Lemos - 201000660

Supervisor : Daniel Castro Silva
Co-Supervisor: Leonardo Ferreira

July 26, 2022

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Master in Informatics and Computing Engineering