



Apprentissage Automatique Project

Bilal SAID

Ph.D. AI, Research Project Manager

SOFTEAM
UNE MARQUE DE DOCAPOSTE


DOCAPOSTE
l'avenir devient plus simple

Outline

- Aim
- Scope
- Teams
- Deliverable
- Deadlines
- Evaluation

Aim

- Apply what we learned on new unseen datasets
 - Data discovery, cleaning, exploration, analysis, split, and transformation
 - Model training, evaluation, comparison, and hyperparameter optimization
 - Analysis of the results, critical thinking and outlook for future enhancement
- Get inspired by others (including ChatGPT, GitHub, Kaggle, friends, our lab codes...)
- Master the content
- Show a coherent methodology
- Justify with clear reasoning

Scope

- Target Datasets

- Small/Medium with 10s of features (10-20) & few 10-100Ks instances
- Dense, mostly clean (well structured, not too many missing values...)
- Previously used and verified (known source, not artificial...)
- Numeric AND Categorical features

362 datasets found

dense

>10 features

>10000 instances

verified



- **Optional & Complementary Only:** pure text (e.g., reviews), images (product photo)...



- **Better Excluded:** simulated/generated datasets, time series, dynamic stream...

- Target ML Tasks

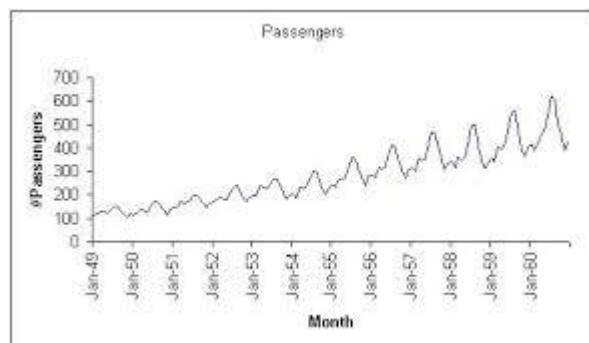
- Regression
- Classification (binary or multi-class)
- Hyper-param tuning



- **Optional & Complementary Only:** deep learning (text, images), LLM...



Dataset	DenseNet- -CT_CAM	FCN- -CT_CAM	SLCNN- -CT_CAM	TopNet	MLSTM- -PCM	WBSS- +MSLE	ED- -INN	DTW- -INN-D	ED-INN (norm)	DTW-INN (norm)	DTW-INN (global)
Antikythera Mechanism	0.9997	0.9997	0.9997	0.987	0.975	0.99	0.97	0.98	0.987	0.97	0.98
Asia	0.4667	0.4667	0.4667	0.333	0.267	0.33	0.267	0.267	0.2	0.267	0.22
Attrition	1	1	1	1	0.81	1	0.675	1	0.875	1	0.875
Bio-Mimics	0.9997	0.991	0.8186	0.997	0.985	0.99	0.964	0.969	0.964	0.969	0.989
Character Trajectories	0.9997	0.991	0.8186	0.997	0.985	0.99	0.964	0.969	0.964	0.969	0.989
Chemical Compound	0.3692	0.359	0.3602	0.356	0.315	0.315	0.319	0.313	0.329	0.319	0.329
Handwritten Digit	0.473	0.4665	0.4459	0.378	0.365	0.365	0.279	0.306	0.231	0.278	0.306
Hepatitis	0.8184	0.8098	0.7196	0.151	0.085	0.717	0.87	0.849	0.717	0.679	0.717
Image Segmentation	0.9997	0.9997	0.9997	0.64	0.59	0.999	0.999	0.999	0.999	0.999	0.999
Image Segmentation Babu	0.9813	0.9869	0.7213	0.959	0.899	0.91	0.85	0.883	0.85	0.883	0.883
IMIS-SF	0.9997	0.7549	0.7746	0.151	0.089	0.914	0.785	0.754	0.711	0.785	0.711
ForeDigits	0.9817	0.9886	0.9196	0.98	0.936	0.948	0.913	0.914	0.917	0.915	0.919
Phishing	0.3388	0.3036	0.3036	0.175	0.11	0.09	0.154	0.151	0.156	0.164	0.151
SelfRegulation	0.5944	0.5944	0.5889	0.58	0.472	0.46	0.485	0.533	0.539	0.485	0.535
SKT	0.9997	0.9997	0.9997	0.99	0.982	0.967	0.96	0.963	0.967	0.959	0.963
Spoken Arabic	0.9813	0.9523	0.8841	0.981	0.99	0.982	0.967	0.96	0.963	0.967	0.963
Stone Wall Lung	0.6667	0.5	0.5	0.4	0.067	0.33	0.2	0.33	0.2	0.2	0.33
Arg. Value	0.746	0.7513	0.7181	0.681	0.631	0.66	0.665	0.679	0.655	0.666	0.658
Wine&Ties	0	2	3	2	0	2	0	1	0	0	1
Avg. Rank	2.067	2.7	4.367	3.633	7.539	3.819	0.5	8.3	7.33	9.7	8.5



Teams

- Two students per team (i.e., 12 teams)
- Fill-in this form <https://forms.gle/JEvfWHzqGcu6kvgK9>

Deliverable

- Single shared Google Colab Notebook
- Regularly pushed on a publicly shared GitHub repository
- Well-documented
- Organized into separate sections
- Minimal extra work to load dataset and install extra python libs

Notebook Structure (1/3)

- Text cell for Project & Dataset Description
 - Project aim(s), existing solutions
 - Dataset original source URL(s)
- Code cell containing **ALL** libs imports
 - Comments for libs versions or installation instructions, if any
- Data Access Instructions
 - Dataset GitHub-hosted project-version URL(s)
 - Data Load Code to access it through GitHub (if slow or technically not feasible, then from G. Drive using a Google Collab)

Notebook Structure (2/3)

Dataset Exploratory Analysis

- Metadata: nb. instances, nb. features, types + your comments
- Nb. nulls or na : code for identification, then your interpretation, & possible and chosen strategies to cope with them with arguments
- Features values distributions, scaling & outliers: possible and chosen strategies to cope with them with arguments
- Target feature study: distribution / class balance, your interpretation & decision to cope with it
- Features correlation, selection, dimension reduction & anomalies detections: your explanation, decision on whether to proceed with different data input in # pipelines
- Unsupervised clustering: correlation with target feature,
- Interpretations, hypotheses and conclusions

Notebook Structure (3/3)

ML Baseline & Ensemble Models

- Training / Validation / Test Splits
- Pipelines & Models
- Training / Validation Code
- Test Code
- Interpretation & Discussion of Results

Enhanced Models, Hyperparameter Tuning & Analysis

- Justification of choices
- Interpretation and analysis of the obtained results

Deadlines

Date	Submission
Nov 14	Teams & Topics Data Exploratory Analysis & Unsupervised Exploration
Nov 28	ML Baseline & Ensemble Models
Jan 23	Enhanced Models, Hyperparameter Tuning & Analysis Final Submission
Feb 03	Defense

Evaluation

[Tentative]

- 5% Following Instructions
- 5% Notebook organization
- 30% Data Exploratory Analysis
Unsupervised Exploration
- 30% ML Baseline & Ensemble Models
- 30% Enhanced Models, Hyperparameter Tuning & Analysis