APC

MD3 Kaggle

Marc.Borras@uab.cat

UAB
Universitat Autònoma
de Barcelona

https://www.kaggle.com/search?q=machine+learning

MD3 Kaggle (**25%** Nota final, optativa, individual i no recuperable)

Explicació didàctica d'un cas pràctic:

- crear un **repositori Github** on s'expliquen els diversos passos realitzats per a la resolució d'un problema d'Aprenentatge Computacional.

Els projectes seran aplicats a bases de dades escollides de la plataforma Kaggle, i constaran de tres parts:

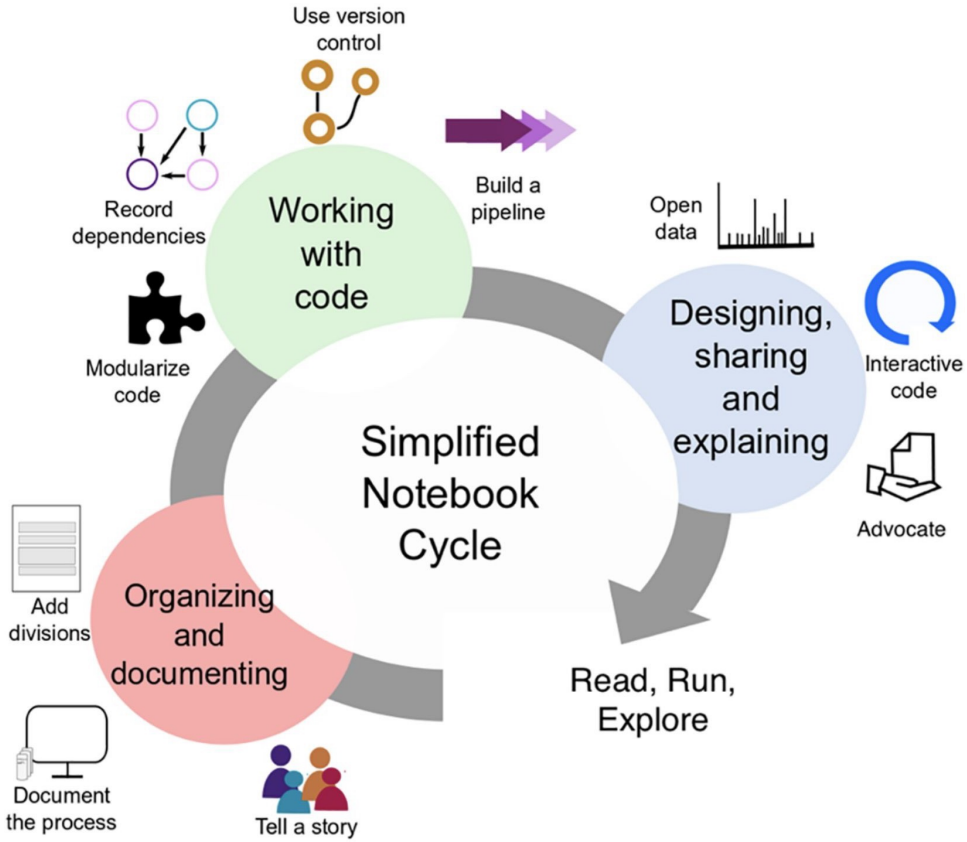
- una explicació dels atributs més importants de la base de dades i de l'atribut a predir/classificar;
- breu descripció del mètode d'aprenentatge computacional aplicat, juntament amb els paràmetres escollits;
- i una presentació dels resultats que s'han obtingut.

Exemples de jupyter notebooks es poden trobar en el següent repositori: https://datauab.github.io/

MD3 Kaggle

Avaluació (25% Nota final):

- 10% justificació del problema (introducció i conclusions)
- 25% anàlisis dels atributs
- 25% aplicació de diversos mètodes d'aprenentatge
- 30% visualització i presentació dels resultats
- 10% estructura del repositori github



Github

Aquest any l'entrega es realitza a github per fomentar cooperació i bones pràctiques.

Es valorarà continuïtat de commits, presentació, claredat, atractiu, autocontingut..

Codi, Markdown explicatiu, Notebooks i link al dataset

https://blogs.uab.cat/ticuab/2020/12/13/github-education-a-labast-de-tots-els-uabers/

https://education.github.com/discount_requests/student_application

Sessió 1: Introducció als Projectes Kaggle, Jupyter

Explicació del projecte Kaggle a Github

Introducció a les eines

- jupyter notebook, lab
- github
- llibreries de ML (numpy, scipy, scikit-learn, pytorch, tensorflow, onnx, xgboost, catboost) https://github.com/ml-tooling/best-of-ml-python
- DevOps - MLOps

Sessió 1: llibreries, github, dades

Objectiu:

- Projecte assignat
- Set-up github, notebook i llibreries
- Entendre dades i tasca

Establir variables objectius i definició de la tasca, variable objectiu?

Seleccionar dades: quins són els atributs importants, per la tasca?

Netejar dades, treure Outliers, Normalitzacions

Sessió 2: data mining

Objectiu:

Processar les dades

Veure patrons - correlacions

PCA vs TSNE: reducció dimensionalitat

Definir conjunts d'aprenentatge (train-val-test)

Aplicar models bàsics scikit-learn

Sessió 2: model learning

Objectiu:

- Predicció dels models per nous exemples (deploy)
- Optimització d'hiperparàmetres
- Comparativa de models

Proves amb models més potents, altres llibreries..

Sessió 3: comparativa de models

Machine Learning Workflow: How to evaluate a model

Sessió 3: anàlisi de resultats

Objectius:

- Entendre resultats
- Mètriques correctes i fiables
- Conclusions a extreure del model i les dades

A partir de les dades, què podem fer per millorar-ne els resultats

Comparativa amb altres treballs

Sessió 4: Github

Treballar el repositori.

Objectius:

- resoldre dubtes
- rebre feedback

Com presentar, organitzar i mantenir el propi repositori de github

Com transmetre confiança del codi. Resultats fiables i reproduïbles?

Sessió 5: Avaluació cas Kaggle

Presentació de 7 min per estudiant amb slides amb link desde github

- 10% justificació del problema (introducció i conclusions)
- 25% anàlisis dels atributs (data mining)
- 25% aplicació de diversos mètodes d'aprenentatge
- 30% visualització i presentació dels resultats
- 10% presentació de l’estat del repositori Github

Sessió 1 de Treball cas Kaggle: llibreries, github, dades

MD3 Kaggle

Sessió de Treball cas Kaggle: llibreries, github, dades

Objectiu:

- Projecte assignat
- Set-up github, notebook i llibreries
- Entendre dades i tasca

Establir variables objectius i definició de la tasca, variable objectiu?

Seleccionar dades: quins són els atributs importants, per la tasca?

Netejar dades, treure Outliers, Normalitzacions

Projecte Assignat

Hem escollit datasets d'entre 500KB i 15MB segons la API de Kaggle i son CSV

Hi ha varis tipus de datasets:

- Tabulars (~70%)
- NLP (~20%) https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
- Visió (~5%)
- Series Temporals (~5%)

Si creieu que n'heu vist un de millor (que us és més interessant, es pot cambiar), però aviseu per a que no hi hagi datasets repetits.

Setup Github + llibreries

Crear nou repositori: https://github.com/new

git clone per descarregar-lo

pip install o conda install per les llibreries

Guies de best practices i tips a tenir en compte:

- https://www.jeremyjordan.me/ml-projects-guide/
- https://cmawer.github.io/reproducible-model/
- https://neptune.ai/blog/how-to-organize-deep-learning-projects-best-practices
- https://towardsdatascience.com/organizing-machine-learning-projects-e4f86f9fdd9c

Exemple repositori

```
├── README.md                    <- You are here
│
├── config                       <- Directory for yaml configuration files for model training, scoring, etc
│   ├── logging/                 <- Configuration of python loggers
│
├── data                         <- Folder that contains data used or generated.
│   ├── archive/                 <- Place to put archive data is no longer used. Not synced with git.
│   ├── external/                <- External data sources, will be synced with git
│   ├── sample/                  <- Sample data used for code development and testing, will be synced with git
│
├── demo                         <- Folder that contains examples on how to use the code for simple executions.
│
├── docs                         <- A default Sphinx project; see sphinx-doc.org for details.
│
├── figures                      <- Generated graphics and figures to be used in reporting.
│
├── models                       <- Trained model objects (TMOs), model predictions, and/or model summaries
│   ├── archive                  <- No longer current models. This directory is included in the .gitignore and is not tracked by git
│
├── notebooks
│   ├── develop                  <- Current notebooks being used in development.
│   ├── archive                  <- Develop notebooks no longer being used.
│
├── src                          <- Source data for the sybil project
│   ├── archive/                 <- No longer current scripts.
│   ├── helpers/                 <- Helper scripts used in main src files
│   ├── ingest_data.py           <- Script for ingesting data from different sources
│   ├── generate_features.py     <- Script for cleaning and transforming data and generating features used for use in training and scoring.
│   ├── train_model.py           <- Script for training machine learning model(s)
│   ├── score_model.py           <- Script for scoring new predictions using a trained model.
│   ├── postprocess.py           <- Script for postprocessing predictions and model results
│   ├── evaluate_model.py        <- Script for evaluating model performance
│
├── test                         <- Files necessary for running model tests (see documentation below)
│   ├── true                     <- Directory containing sources of truth for what results produced in each test should look like
│   ├── test                     <- Directory where artifacts and results of tests are saved to be compared to the sources of truth.
│   ├── test.py                  <- Runs the tests
│
├── run.py                       <- Simplifies the execution of one or more of the src scripts
├── requirements.txt             <- Python package dependencies
```

Sessió 2 de Treball cas Kaggle: data mining i model learning

MD3 Kaggle

Sessió de Treball cas Kaggle: data mining

Objectiu:

Processar les dades

Veure patrons - correlacions

PCA vs TSNE: reducció dimensionalitat

Definir conjunts d'aprenentatge (train-val-test)

Aplicar models bàsics scikit-learn

Processat de dades

Info extreta de https://www.cs.uoi.gr/~tsap/teaching/2012f-cs059/slides-en.html

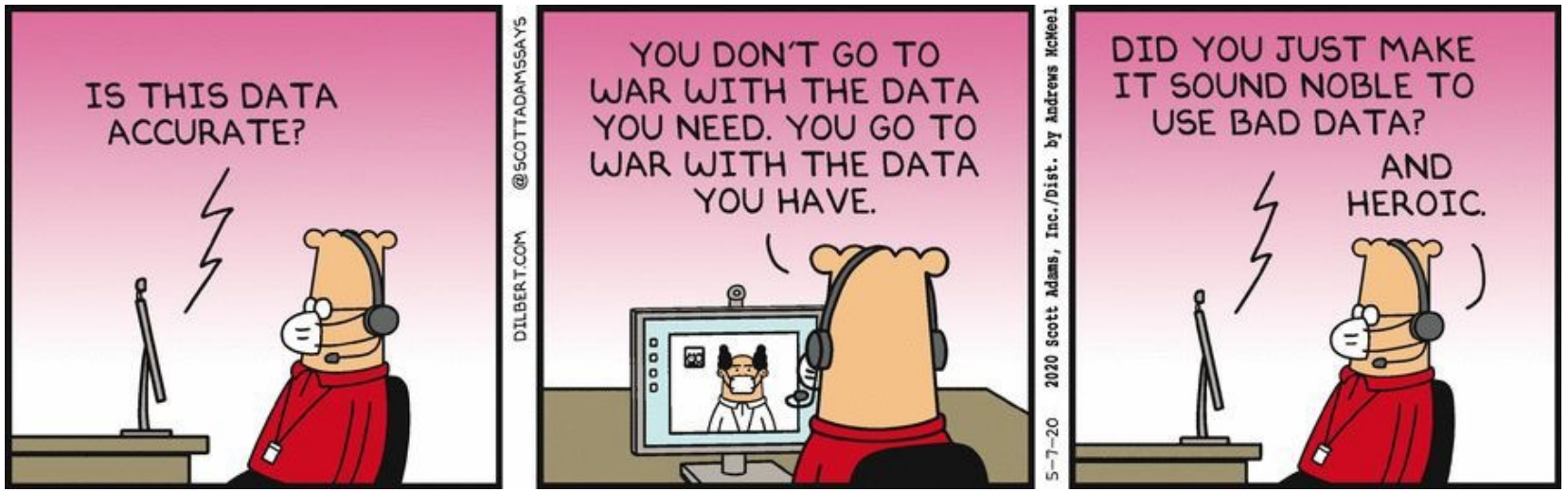
DATA MINING
LECTURE 2

Data Preprocessing
Exploratory Analysis
Post-processing

Visualització dades i patrons

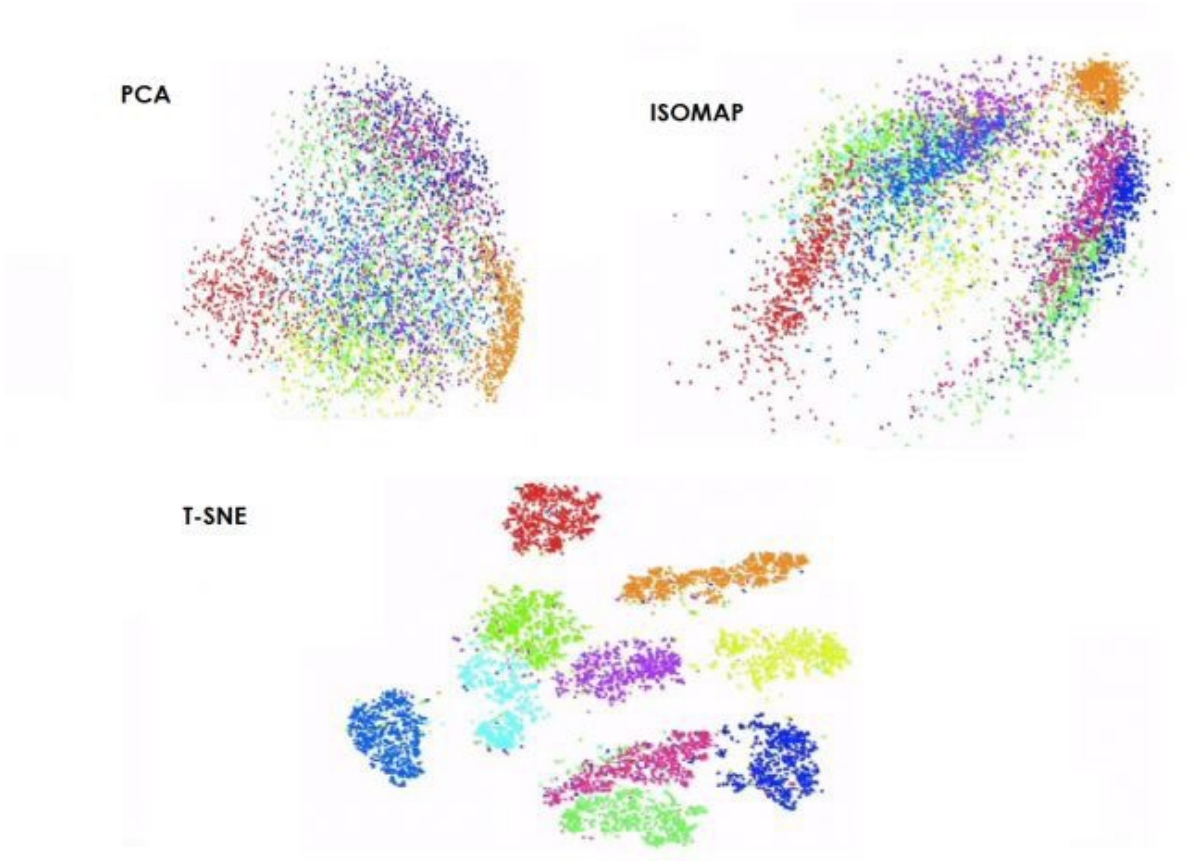
Veure dades amb seaborn:

- https://towardsdatascience.com/14-data-visualization-plots-of-seaborn-14a7bdd16cd7



Processat de dades

PCA - T-SNE



Sessió de Treball cas Kaggle: model learning

Objectiu:

- Predicció dels models per nous exemples (deploy)
- Optimització d'hiperparàmetres
- Comparativa de models

Proves amb models més potents, altres llibreries..

Entrenament de models de AI

Teniu la variable objectiu, enteneu el problema?

Teniu les dades separades en subconjunts de train-val-test? o bé train-test?

Quin model heu escollit? per què?

Quina mètrica?

extra:

/**demo**: Predicció dels models per nous exemples (deploy)

codi que carregui model, i l'apliqui sobre unes dades noves. Genereu-les o busqueu-ne

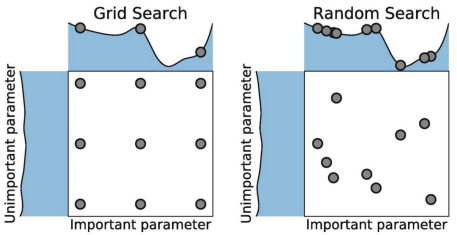
Optimització d'hiperparàmetres

Què son?

Com es diferencia dels paràmetres?

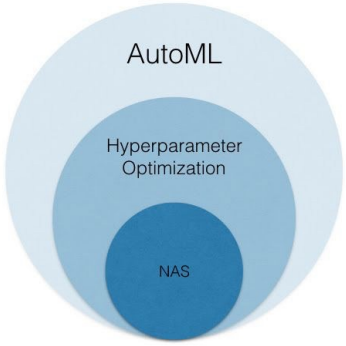
Sabeu quin mètode d'optimització fa servir el vostre model?

Com afecten al teu model?



De quina forma pots trobar els millors valors? Trial and Error, Grid Search, Random Search, Bayesian Optimization.

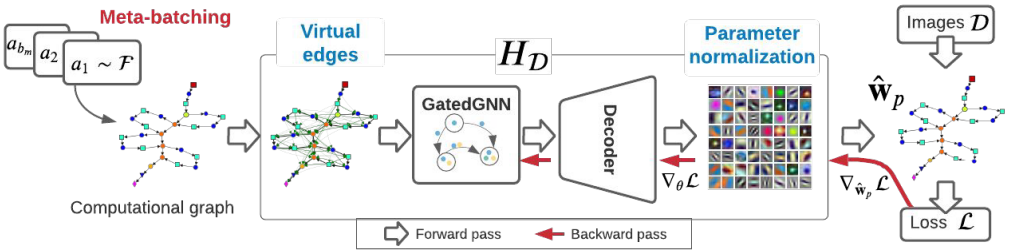
https://medium.com/analytics-vidhya/hyperparameter-search-part-1-2b67fd7a71d8



Parametres vs hiperparàmetres:

Parameter Prediction for Unseen Deep Architectures (NeurIPS 2021)

authors: Boris Knyazev, Michal Drozdzal, Graham Taylor, Adriana Romero-Soriano



Sessió 3 de Treball cas Kaggle: anàlisi de resultats

MD3 Kaggle

Sessió de Treball cas Kaggle: anàlisi de resultats

Objectius:

- Entendre resultats
- Mètriques correctes i fiables
- Conclusions a extreure del model i les dades

A partir de les dades, què podem fer per millorar-ne els resultats

Comparativa amb altres treballs

Sessió de Treball cas Kaggle: anàlisi de resultats

- Quins models heu provat?
- Quins hiperparàmetres?
- Diferents pre-processats?
- Podeu comparar eficiència / rendiment de models

Proves amb models més potents, altres llibreries..

- Scikit: Logistic Regression, Classifier, SVM, Random Forest…
- XGBoost: Gradient Boosting
- CatBoost
- Pytorch

Comenceu a pensar com es comporta el model (amb les mètriques) i com seria millor continuar per millorar els resultats

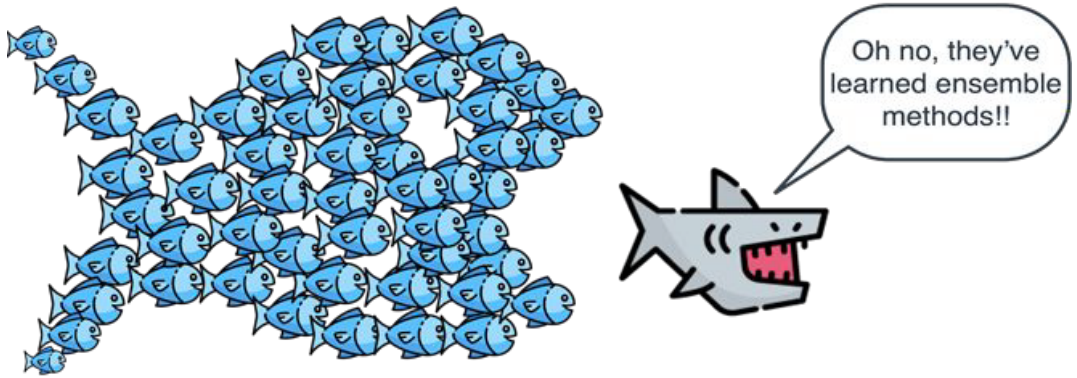
Sessió de Treball cas Kaggle: anàlisi de resultats

Entendre resultats:

- Quina mètrica utilitzeu? Sabeu què significa?
- Conjunts independents d'aprenentatge-validació?
- Analitzar on falla el model.
    - Mostrar exemples bons i dolents. Hi trobeu alguna explicació?
- Podeu fer servir el coneixement après per modificar el algorisme (preprocessat, model, dades..) i que finalment funcioni millor

Sessió de Treball cas Kaggle: anàlisi de resultats

Millora de models (ensembles)

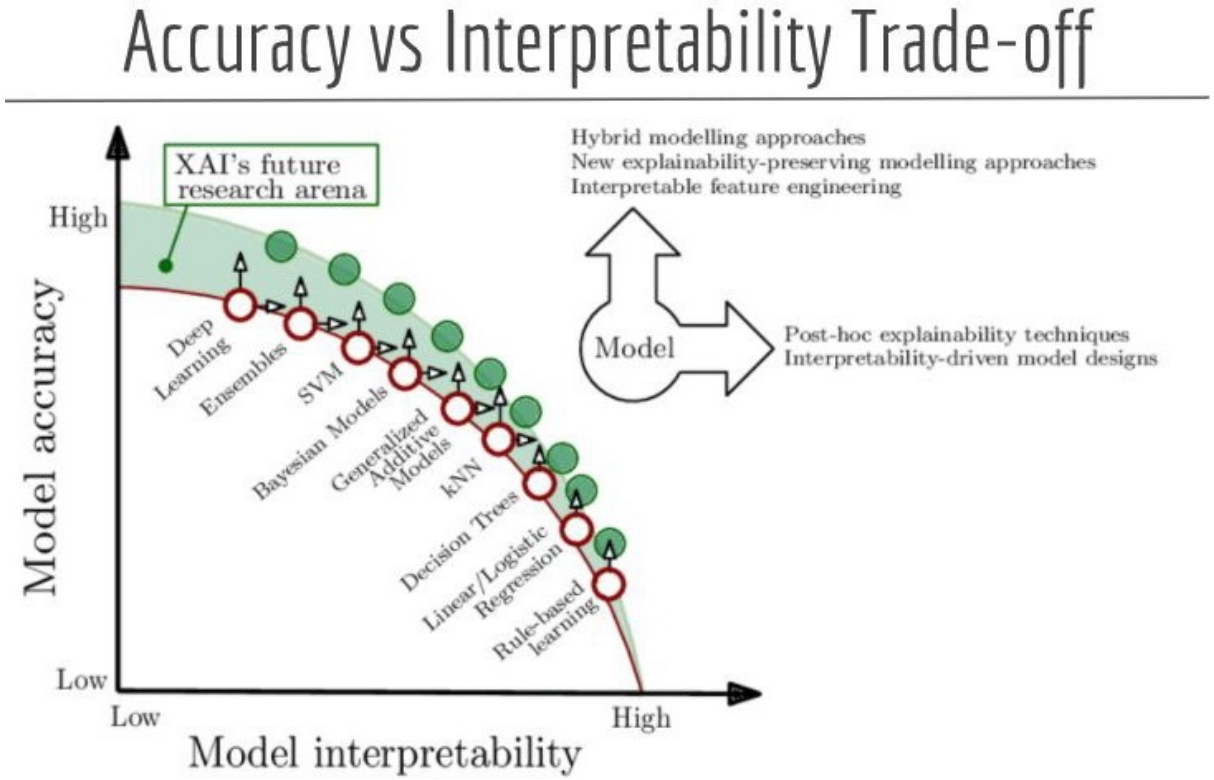


- Heu provat diferents models?
- Tenen diferents resultats?
- Es poden combinar els models?
- Creieu que millorarà la mètrica final?
- En quins casos podeu aconseguir millores?

Sessió de Treball cas Kaggle: anàlisi de resultats

Entendre els models

- Sabeu interpretar el perquè de una predicció??
- Creieu que és important?
- Interpretability, Explainability & Auditability



https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools

Sessió de Treball cas Kaggle: anàlisi de resultats

- En mode de deploy (quan el model està en producció):
  - Si es una classificació, com escolliries si la predicció és confiable?
  - Com seleccionar el threshold?
- Podeu fer codi de demostració? Dona't un exemple, com aplicar el model
- Github:
  - Teniu exemples de codi al github de funcionament del model?
  - Teniu taules de resultats del model?
  - Us compareu amb codis d'altres desenvolupadors?

Sessió 4 de Treball cas Kaggle: Github

MD3 Kaggle

Sessió de Treball cas Kaggle: Github

Treballar el repositori

Objectius:

- resoldre dubtes
- rebre feedback

Com presentar, organitzar i mantenir el propi repositori de github

Com transmetre confiança del codi. Resultats fiables i reproduïbles?

Exemple repositori

```
├── README.md                    <- You are here
│
├── config                       <- Directory for yaml configuration files for model training, scoring, etc
│   ├── logging/                 <- Configuration of python loggers
│
├── data                         <- Folder that contains data used or generated.
│   ├── archive/                 <- Place to put archive data is no longer used. Not synced with git.
│   ├── external/                <- External data sources, will be synced with git
│   ├── sample/                  <- Sample data used for code development and testing, will be synced with git
│
├── demo                         <- Folder that contains examples on how to use the code for simple executions.
│
├── docs                         <- A default Sphinx project; see sphinx-doc.org for details.
│
├── figures                      <- Generated graphics and figures to be used in reporting.
│
├── models                       <- Trained model objects (TMOs), model predictions, and/or model summaries
│   ├── archive                  <- No longer current models. This directory is included in the .gitignore and is not tracked by git
│
├── notebooks
│   ├── develop                  <- Current notebooks being used in development.
│   ├── archive                  <- Develop notebooks no longer being used.
│
├── src                          <- Source data for the sybil project
│   ├── archive/                 <- No longer current scripts.
│   ├── helpers/                 <- Helper scripts used in main src files
│   ├── ingest_data.py           <- Script for ingesting data from different sources
│   ├── generate_features.py     <- Script for cleaning and transforming data and generating features used for use in training and scoring.
│   ├── train_model.py           <- Script for training machine learning model(s)
│   ├── score_model.py           <- Script for scoring new predictions using a trained model.
│   ├── postprocess.py           <- Script for postprocessing predictions and model results
│   ├── evaluate_model.py        <- Script for evaluating model performance
│
├── test                         <- Files necessary for running model tests (see documentation below)
│   ├── true                     <- Directory containing sources of truth for what results produced in each test should look like
│   ├── test                     <- Directory where artifacts and results of tests are saved to be compared to the sources of truth.
│   ├── test.py                  <- Runs the tests
│
├── run.py                       <- Simplifies the execution of one or more of the src scripts
├── requirements.txt             <- Python package dependencies
```

Exemple repositori mínim

```
├── README.md                   <- You are here
│
├── data
│   ├── external/
│
├── demo
│
├── models
│
├── notebooks
│   ├── develop
│
├── src
│   ├── generate_features.py
│   ├── train_model.py
│   ├── score_model.py
│
├── requirements.txt
```

<- Folder that contains data used or generated.
<- External data sources, will be synced with git

<- Folder that contains examples on how to use the code for simple executions.

<- Trained model objects, model predictions, and/or model summaries

<- Current notebooks being used in development.

<- Source data for the sybil project
<- Script for cleaning and transforming data.
<- Script for training machine learning model(s)
<- Script for scoring new predictions using a trained model.

<- Python package dependencies

Si poseu notebooks, si son executats molt millor!

Si les dades son grans (> 5MB), no cal que les pujeu al github.

Exemple de

Readme.md

Pràctica Kaggle APC UAB 2022-23
Nom: *****
DATASET: *****
URL: [kaggle](http://....)

Resum
El dataset utilitza dades de...
Tenim X dades amb N atributs. Un % d'ells és categoric / els altres són numérics i estàn normalitzats...

Objectius del dataset
Volem aprender quina és la ...

Experiments
Durant aquesta pràctica hem realitzat diferents experiments.

Preprocessat
Quines proves hem realitzat que tinguin a veure amb el pre-processat? com han afectat als resultats?

Model
Model	Hiperparametres	Mètrica	Temps
[Random Forest](link)	100 Trees, XX	57%	100ms
Random Forest	1000 Trees, XX	58%	1000ms
SVM	kernel: lineal C:10	58%	200ms
[model de XXX](link al kaggle)	XXX	58%	?ms
[model de XXX](link al kaggle)	XXX	62%	?ms

Demo
Per tal de fer una prova, es pot fer servir amb la següent comanda
``` python3 demo/demo.py --input here ```

Conclusions
El millor model que s'ha aconseguit ha estat...
En comparació amb l'estat de l'art i els altres treballs que hem analitzat....

Idees per treballar en un futur
Crec que seria interesant indagar més en...

Llicencia
El projecte s'ha desenvolupat sota llicència ZZZz.
```

Sessió 5 d'Avaluació

MD3 Kaggle

Avaluació cas Kaggle

Presentació de 7 min per estudiant amb slides amb link desde github

- 10% justificació del problema (introducció i conclusions)
- 25% anàlisis dels atributs (data mining)
- 25% aplicació de diversos mètodes d'aprenentatge
- 30% visualització i presentació dels resultats
- 10% presentació de l’estat del repositori Github

Avaluació cas Kaggle

10% justificació del problema (introducció i conclusions)

- Explicacions i referencies (5%)
- Conclusions i future work (5%)

25% anàlisis dels atributs (data mining)

- Anàlisis Exploratori de les dades (10%)
- Explicació i validació del preprocessat (15%)

25% aplicació de diversos mètodes d'aprenentatge

- Mètodes utilitzats (Dificultat, variabilitat..) (15%)
- Optimització d'hiperparàmetres (10%)

30% visualització i presentació dels resultats

- Taula / Gràfiques resultats (10%)
- Insights de dades. Què podeu extreure de les dades treballades? (10%)
- exemples ben classificats i dolents (10%)

10% presentació de l'estat del repositori Github

- Estat del readme, models, demo, carpetes (5%)
- Evolució continuada del curs en forma de commits (5%)

10% Extra per dificultat, ingeniós i treball realitzat. (No superarà el Màxim del 100%)

Què s'ha d'entregar (Abans de dimecres 14 de desembre 23:55h)

Al caronte:

- zip amb pdf de les slides

A les slides (tindreu **7 minuts** per presentar-les):

- Portada: **Nom, Dataset, Link al github**
- Introducció al dataset / problema que voleu resoldre (1/2m)
- Preprocessat (1m)
- Metodes utilitzats (2m)
- Resultats i exemples (3m)
- Conclusió i Treball a futur (1/2m)

  Podeu posar les slides que volgueu, la restricció és de temps

Al github:

- Que sigui public