Liste de connaissances en sciences de données

Codes Kaggle : https://github.com/pbranchini/Kaggle_codes

Feature engineering

Data preprocessing
- Data scaling
- Fill missing values
- Categorical/Label encoding
- Frequency encoding

Target encoding
- Mean/sum/difference encoding
- Regularization techniques

Text features
- Bag of words
- Tf-idf
- Ngrams
- GloVe
- Word2Vec
- Stemming/Lemmatization

KNN features

Supervised Learning

Linear models
- Linear regression
- Polynomial regression
- Lasso/Ridge/ElasticNet regressions
- Logistic/Softmax Regression
- Perceptron

Decision Trees
- Boosted trees
- Extra trees
- Random Forest
- Extra trees

Support Vector Machines (SVM)
- SVC and SVR

Nearest Neighbors (KNN)

Neural Networks
- Multilayer perceptron
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)
- Long Short-Term Memory Networks (LSTM)
- Gated Recurrent Unit (GRU)
- Residual Networks (ResNet)
- Inception Networks

Unsupervised Learning

Principal Component Analysis (PCA)

Gaussian Mixture model

K-Means algorithm

TSNE algorithm

Autoencoders

Model selection & evaluation

Hyperparameters tuning
- Grid search & randomized search

Regression metrics
- RMS(L)E, MA(P)E, R^2

Classification metrics
- Logloss
- Accuracy
- AUC (ROC curve)
- Cohen's Kappa
- Hinge loss
- Confusion matrix
- Precision & Recall

Metrics optimization

Cross-Validation for model evaluation
- Holdout
- KFold
- Leave-one-out (LOO)
- CV for time series

Ensemble methods

Voting Classifier

Adaboost algorithm

Gradient boosting

Stacking

Time Series models

(S)ARIMA,

(N)GARCH

Vector Autoregressive (VAR)

Other

Data split techniques
- Random
- Timewise
- By IDs

Optimization algorithm
- Stochastic gradient descent
- RMS Prop
- Adam Optimization

Programming languages
- Python
- Matlab
- R

Python libraries

Pandas, Numpy, Matplotlib, Seaborn, Tensorflow, Keras, Sklearn, XGboost, Catboost, Lightgbm