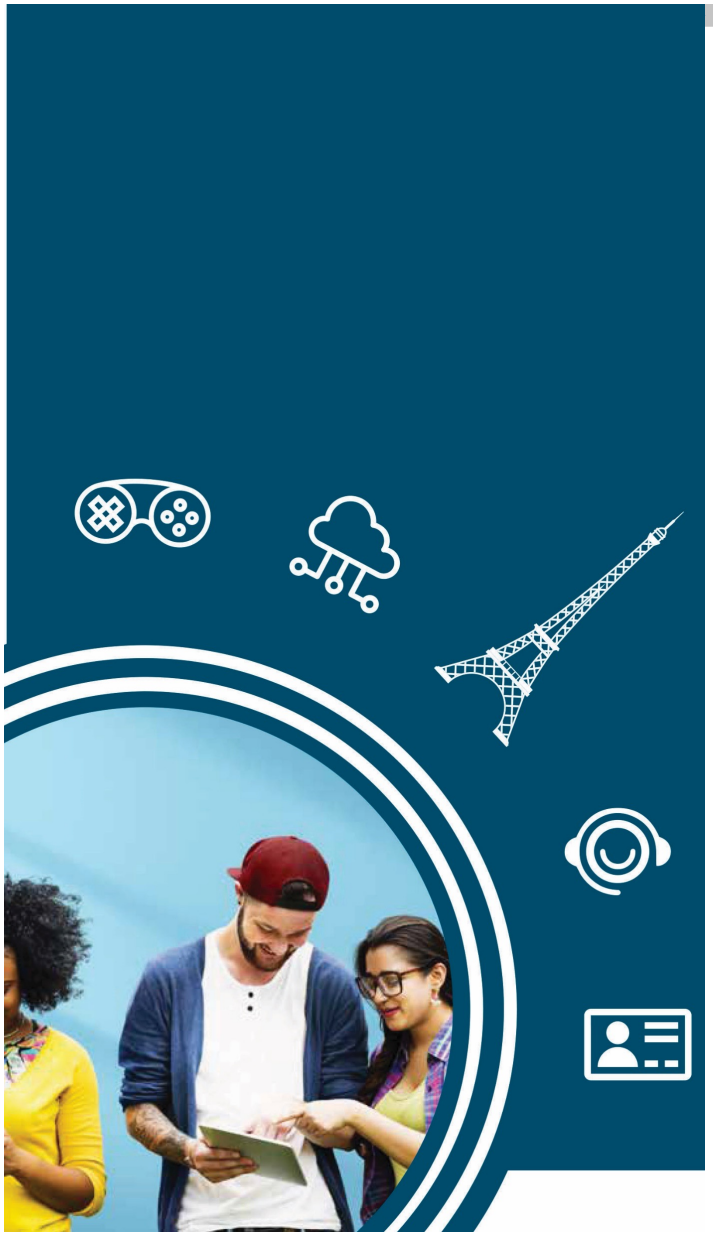
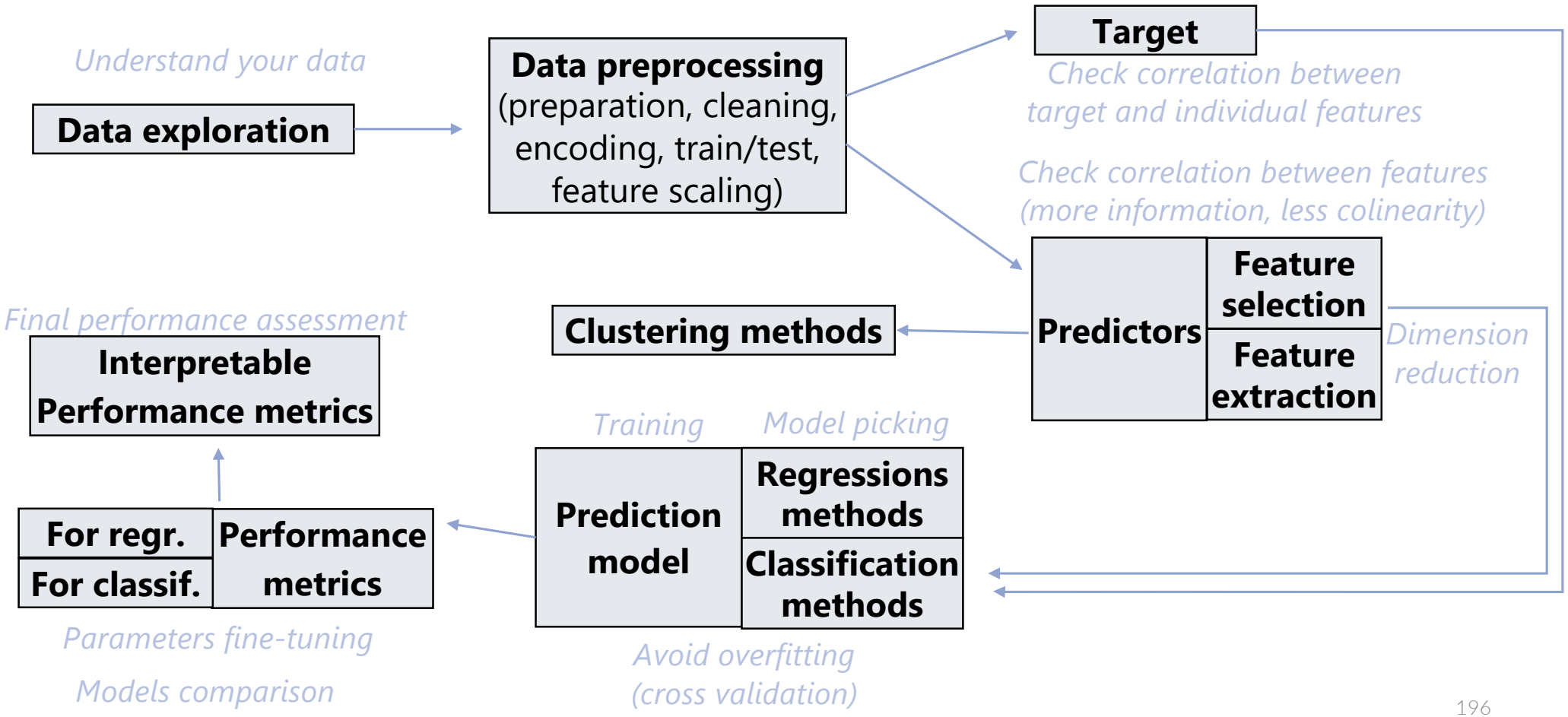


EXTENSIONS

GENERAL ML APPROACH



METHODS OVERVIEW

Regression

Algorithm	Advantages	Disadvantages
Linear regression	Works on any size of dataset. Easy and simple implementation. Fast training. Value of coefficients gives an assumption of feature significance (interpretablity).	Applicable only if the solution is linear. Algorithm assumes input features to be mutually independent. Input residuals are assumed to be normally distributed.
Polynomial regression	Works on any size of data set. Can handle a broad range of non-linear problems. Fast training. Value of coefficients gives an assumption of feature significance (interpretability).	The right polynomial degree has to be chosen. High sensitivity to outliers. Input residuals are assumed to be normally distributed.

METHODS OVERVIEW

Classification

Algorithm	Advantages	Disadvantages
Logistic regression	Fast to train and predict. Good for small classification data problems. Easy to understand.	Not very accurate. Not suitable for non-linear data. Not flexible to adapt to complex data. Model occasionaly prone to overfitting.
k-Nearest Neighbours	Simple, adaptable to the problem. Accurate. Easy to understand.	Memory intensive since all training data might be involved in decision making. Choosing wrong distance measure can produce inaccurate results.
Decision Trees & Random Forest	High accuracy (especially Random Forest). Good starting point to solve a problem. Flexible and can fit well a variety of different data. Fast to execute. Easy to use.	Slow at training. Risk of overfitting (esp. Decision Trees). Not suitable for small samples. Small change in training data changes model. Occasionally too simple for very complex problems.
Support Vector Machine	Accurate in high-dimensional spaces. Kernel trick can help to solve complex problems. Memory efficient.	Prone to overfitting. No built-in probability estimation. Sensitivity to dataset size.

METHODS OVERVIEW

Clustering

Algorithm	Advantages	Disadvantages
k-Means	Faster than hierarchical clustering. Works very well when k is well chosen. Clusters can be computed in batches to improve performance. Good when all clusters have equal size.	Does not work well with categorical values, when clusters overlap or with clusters of different density. Can be slow when the number of clusters is unknown or large. Potentially returns different clusters each time due to random initialization.
Hierarchical Clustering	Does not require to specify the number of clusters. Easy to implement. Produces a dendrogram, which helps understand the data. Always returns the same clusters.	Can be slow to compute on large datasets. Sometimes difficult to identify the number of clusters with the dendrogram.