Python Machine Learning Cheat Sheet

Machine Learning - Ingeniería de Datos I

4º IITV - 3º ISW - 4º CVAD

Academic year 2025/2026

Antonio M. Durán Rosal

Python Machine Learning Cheat Sheet

Data Preparation

Descriptive Statistics

- **Data loading**: https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html
- **Counting**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.shape.html
- **Mean**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.mean.html
- **Standard deviation**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.std.html
- **Max**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.max.html
- **Min**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.min.html
- **Quantiles**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.quantile.html
- **Summary**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html
- **Missing values**: https://pandas.pydata.org/docs/user_guide/missing_data.html
- **Correlation**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html
- **Skewness**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.skew.html

Data Visualisation

- **Histograms**: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html
- **Density plots**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.html
- **Boxplots**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.html
- **Correlation matrix graph**: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.matshow.html
- **Dispersion matrix graph**: https://pandas.pydata.org/docs/reference/api/pandas.plotting.scatter_matrix.html

Preprocessing

- **Filter methods**: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
- **Wrapper methods**: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html
- **Determine NaN values**:
  - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.isnull.html
  - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.isna.html
- **Remove missing values**: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.dropna.html
- **Univariate imputation**:
  - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.fillna.html
  - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

- **Multivariate imputation**: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html
- **Binarisation**: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
- **Data scaling**: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
- **Data normalisation**: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
- **SMOTE**: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

Clustering

Metrics

- **Davies-Bouldin**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html
- **Silhouette**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
- **Calinski Harabasz**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html
- **Rand Index**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.rand_score.html
- **Adjusted RI**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html
- **Mutual Information**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual_info_score.html
- **Adjusted MI**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_mutual_info_score.html

Algorithms

- **KMeans**: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- **Agglomerative clustering**: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
- **DBSCAN**: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Dimensionality Reduction

Methods

- **PCA**: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Introduction to Supervised Learning

Validation Techniques

- **Holdout**: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
- **KFold**: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
- **LeaveOneOut**: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneOut.html

Regression Metrics

- **MAE**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html

- **MSE**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html
- **R^2**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

Classification Metrics

- **Confusion matrix**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
- **CCR**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
- **Recall**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html
- **Precision**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html
- **F1-Score**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
- **Kappa**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html
- **Brier**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.brier_score_loss.html
- **AUC**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html
- **Report**: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

Hyperparameter Tuning

- **Grid search**: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- **Random search**: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Simple Classifiers and Regressors

Classifiers

- **ZeroR** → `DummyClassifier` with `strategy='most_frequent'`: `https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html`
- **OneR**: `https://rasbt.github.io/mlxtend/user_guide/classifier/OneRClassifier/`
- **KNN**: `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`

Regressors

- **ZeroR** → `DummyRegressor` with `strategy='mean'`: `https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyRegressor.html`
- **KNN**: `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html`

Linear and Logistic Regression

Models

- **Linear regression**: `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html`
- **Logistic regression**: `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

Decision Trees

Models

- **Classification**: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- **Regression**: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

Artificial Neural Networks

Models

- **Classification**: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
- **Regression**: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

Libraries

- **numpy**: https://numpy.org/
- **pandas**: https://pandas.pydata.org/
- **matplotlib**: https://matplotlib.org/
- **scikit-learn**: https://scikit-learn.org/stable/
- **imbalanced-learn**: https://imbalanced-learn.org/stable/
- **MLxtend**: https://rasbt.github.io/mlxtend/